SOCIAL SENSING GAMES

BY

JUAN FERNANDO MANCILLA CACERES

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Eyal Amir, Chair, Director of Research
Professor Dorothy Espelage Dorothy Espelage
Associate Professor Roxana Girju
Associate Professor Karrie Karahalios
Principal Research Scientist Henry Lieberman, MIT Media Lab

ABSTRACT

We introduce *Social Sensing Games* (SSGs), a new method for collecting data about social relationships, and present an algorithm that can be used to efficiently infer information from the output of the games. The main purpose of this new method is to use people's online interactions to learn about their offline behavior.

Traditionally, scientists have studied social interactions through the use social networks obtained through carefully designed self-report surveys that impose limitations on the types of research questions that can be answered. Recently, thanks to the ever-increasing use of computers and mobile devices for managing social relationships, scientists look to use large amounts of data that is easily accessible. Unfortunately, this latter data lacks the experimental and theoretical validity of previous methods.

SSGs address these concerns by providing an interface that can collect fine-grained data that is relevant to the research question at hand. The first contribution of this thesis is the formalization of Social Sensing Games in such a way that they combine the power of lab-controlled experiments, the detailed observations of observational studies, and the scalability and inferential power of computational methods. This new definition can be used by researchers to easily design SSGs specific to the problem they wish to address.

The second contribution is most relevant to the field of social network analysis. We present an algorithm for analyzing the output of SSGs which main insight is that, in some cases, pairwise relationships are enough to infer global attributes of the nodes encoded in a social network and that such assumption may reduce the complexity of inference, help with the scarcity of data, and still maintain some of the context of the network.

We show these contributions through two applications: The evaluation of commonsense knowledge, and the identification of classroom aggressors (or *bullying*). The second application being in itself an important contribution that provides new insights concerning the study of bullying and cyberbullying.

To my lovely wife Sara and my inspiring parents and siblings.

ACKNOWLEDGMENTS

Every successful journey is the result of hundreds of blessings in the form of help and support from expected and unexpected sources. For all of them, I am eternally grateful to God for providing me with the aid, support, and friendship of the many people that helped me get here and complete this thesis.

First of all, I would like to thank my advisor Eyal Amir. From the moment that I first stepped into his office, he has been a helpful source of advice and ideas. I truly appreciate the opportunity that he gave me by welcoming me into his group and allowing me to pursue my personal research interests. He taught me most of what I know about Artificial Intelligence and showed me how to be an independent researcher, to trust my ideas, and to always aim higher. He was extremely helpful during my PhD studies and I will always be grateful. Thanks Eyal!

I would also like to thank Dorothy Espelage. Not only was she the one that provided the original direction of my research, but she was always helpful and willing to meet and discuss my work. Thank you very much Dorothy for allowing me to work with you and explore such an interesting research topic.

I also wish to thank Henry Lieberman for all his helpful feedback and insight into my work. It is a unique opportunity to be able to work with someone that shares so many research interests with me. All the conversations that we had influenced my work, and I appreciate his willingness to be a member of my thesis committee. I also want to thank Karrie Karahalios for giving me the chance to explore alternative approaches and applications of my research, and for showing me showed how Computer Science can be applied to different intellectual areas. Last but not least, I want to thank Roxana Girju for all her comments on my work and her feedback as part of my thesis committee. Her contributions helped me do a much better job.

I also want to thank all the current and past members of the General Intelligence Group (GIG), previously known as the Knowledge Representation and Reasoning (KRR) Group: Deepak Ramachandran, Afsaneh Shirazi, Hannaneh Hajishirzi, Jaesik Choi, Mark Richards,

Abner Guzman, Wen Pu, and Codruta Girlea. Through their comments and exchange of ideas they have helped me to become the researcher I am today. I am glad to have met such a diverse group of intelligent people and certainly wish that our paths may cross again.

My life and experiences during my graduate studies would not have been the same without the many people with whom I shared ideas and time. I am particularly thankful for the friendship and collaboration with Leonardo Bobadilla and Oscar Sanchez. They were my partners from the beginning of my studies and helped me in every step of the way. I am doubtful that I would have been as successful without our Saturday Qual/Research reading group. I appreciate all the ideas, jokes, and time that we shared. Thank you very much.

Also, I would like to thank the many other friends that helped me at different moments of my life. Among them Pedro Crisostomo and Esteban Meneses, my Latin-American Fulbright friends, Thyago Duque and Minas Charalambrides for their help in making my stay in Urbana much more interesting and fun, and Peter Dinges for all the conversations that we had and for sharing Shinkendo with me, *arigatou gozaimashita!* I also want to thank Yonatan Bisk, Daphne Tsatsoulis, Tarun Prabhu, and Cem Subakan for their help and feedback during my preparation for the final defense of this thesis.

Finally, I would like to thank my family, starting with my parents, Carlos Mansilla M. and Carmen Cáceres de Mancilla, for they have always been my role models and standard to which I compare myself and others. I could not have wished for better parents, they've been the example of tenacity and morality through all my life, and with their example and love they have given me an almost unfair advantage against the whole world. *Muchas gracias por todo Papa y Mama, este logro es también de ustedes!*

Also, I want to thank my brothers and sister, Carlos, Tatiana and Pablo. Thank you very much for all the help and encouragement that you have given me throughout my life. You all have been as second parents to me and I am glad to call you my family. Your advices and experiences have all serve as inspiration to me in different aspects of my life.

And finally, I want to give a special thank you to my wife, Sara Estrada Villalta, without whose support, inspiration, and help, my PhD studies would not have been possible. Our journey together started long before my graduate studies and without her encouragement I would probably not have even started (she was the one that suggested to apply to UIUC!). She has served both as muse and colleague, and has probably read and corrected all my papers, and without her input, I doubt they would have the impact they did. All in all, this success is equally hers. Sara, as the song says "you are a full-time lover, and a full-time RA". Thank you very much for sharing your life with me.

I wish to acknowledge the support of the Fulbright/LASPAU scholarship program that helped me start my studies. I was also partially supported by the Beckman Institute and University of Illinois Cognitive Science / Artificial Intelligence Research Award CS/AI 2012, National Science Foundation (NSF) IIS grant 09-17123-RI: Scaling Up Inference in Dynamic Systems with Logical Structure, and NSF IIS grant 09-68552-SoCS: Analyzing Partially Observable Computer-Adolescent Networks. I also want to give thanks to SBP-12, AAAI-12, SBP-13, and IUI-13 for granting me travel scholarships.

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION 1
1.1 Collecting Data About Social Interactions 2
1.2 Analyzing Social Interactions 3
1.3 Proposed Approach 4
1.4 Structure of the Thesis 6

CHAPTER 2 BACKGROUND 7
2.1 Related Fields 7
2.2 Inference in Social Networks Analysis 18
2.3 About the Applications of Social Sensing Games 22

CHAPTER 3 SOCIAL SENSING THROUGH GAMES 28
3.1 General Notions 28
3.2 Definition of Social Sensor 30
3.3 Definition of Social Game 31
3.4 Definition of Social Sensing Game 32
3.5 Application: Evaluating Commonsense Knowledge 34
3.6 Application: Identification of Aggressive Individuals in Classrooms 45
3.7 Conclusions and Future Work 55

CHAPTER 4 INFERENCE FOR SOCIAL SENSING GAMES 57
4.1 Motivation 57
4.2 General Notions 58
4.3 Global Inference from Pairwise Interactions 60
4.4 Limitations and Heuristics 63
4.5 Application: Efficient Identification of Aggressive Individuals 69

CHAPTER 5 ALTERNATIVE ANALYSIS FOR SOCIAL SENSING GAMES . . . 79
5.1 Visualization of Pairwise Interactions 79
5.2 Understanding Popularity and Computer-Mediated Communication 84
5.3 Analysis of Proactiveness in Aggression 92

CHAPTER 6 PRIVACY AND ETHICAL CONCERNS OF SOCIAL SENSING GAMES 94
6.1 Ethical Risks and Privacy Concerns 94
6.2 Security Measures 97
6.3 Security Measures in Similar Systems 98
6.4 Conclusions 100

CHAPTER 7 RELATED WORK 101
7.1 Visualizations 101
7.2 Work Related to Evaluating Commonsense Knowledge 103
7.3 Work Related to Identifying Aggressive Individuals 104

CHAPTER 8 CONCLUSIONS 106
8.1 Summary of Results 106
8.2 Conclusions and Contributions 107
8.3 Future Directions 107

REFERENCES 110

CHAPTER 1

INTRODUCTION

This thesis introduces Social Sensing Games (SSGs), a new framework to collect data about social interactions and to learn about the attributes of the interacting entities. In particular, SSGs contribute by: 1) introducing a new way of collecting relevant data about social interactions, 2) present an inference algorithm to reason about offline behavior from online interactions, and 3) validate the data and results through grounded behavioral theories.

Social relationships and peer interactions are two of the key factors influencing many social behaviors, and can be studied to further understand and address research questions pertaining to social health (e.g., bullying, smoking, and delinquency), as well as other questions related to social behaviors (e.g, discovering the kind of knowledge that is shared across communities).

These research questions are particularly relevant today due to the increasing use of computers and mobile devices to manage social relationships. In addition, this new availability of data has open new problems and opportunities for both Social and Computer Scientists.

On the one hand, platforms such as online social networks are providing large amounts of data about people and their interactions on a scale previously unimaginable. On the other, social scientists have conducted vast amounts of research analyzing peer relationships, but currently lack the tools to study fine-grained interactions (available through computer-mediated environments) without relying on costly, non-scalable, observational studies.

The contributions here presented are relevant to those interested in areas such as Computational Social Science (CSS), Social Network Analysis (SNA), Artificial Intelligence (AI), and Human-Computer Interactions (HCI). We also present two applications of SSGs. The first one gathers and evaluates commonsense knowledge and is relevant to the Commonsense and AI communities. The second one, focuses on the identification of aggressive individuals in middle-school classrooms (i.e., *bullying*) and is relevant to HCI and Educational Psychology.

1.1 Collecting Data About Social Interactions

The large amounts of data that are currently being generated in daily activities such as shopping, traveling, instant-messaging, etc., allow the measuring of human behavior in ways that were considered impossible just a decade ago. This new availability of data requires new computational tools that are both scalable and capable of using the information to answer important questions.

These challenges are commonly addressed through the combination of techniques developed within Computer Science, Statistics, and the Social Sciences [Lazer et al., 2009]. Common research topics in this area (commonly referred to as Computational Social Science [Cioffi-Revilla, 2010]) include topics such as predicting consumer behavior [Goel et al., 2010] and the spread of influence in social networks [Aral and Walker, 2012], which are studied through the use of large-scale demographics and network data, together with techniques from Agent-Based Modeling (ABM), Machine Learning, and Social Network Analysis.

One of the challenges in this emerging field is that most of the available data has been generated for purposes other than the problems that scientists care about. This means that in many of the cases the data is inappropriate for the problem at hand or the results are difficult to validate.

As an example, assume that we want to explore the relationship between classroom aggression (i.e., *bullying*) and online behavior. Using the available comments from websites like YouTube[1] might not provide appropriate results. The reason being the sense of disembodiment present in online social media [Suler, 2004] that obscures the relationship between the online and offline behavior.

The same argument can be made about data coming from other online sources such as Massive Multiplayer Online games, Online Social Networks (such as Facebook) and others. This does not mean that such research (i.e., studying the dynamics in online environments) is not useful, simply that those dataset are not necessarily appropriate to answer questions about offline behavior.

What it does mean is that, despite the large amounts of data available, relevant useful data for particular social problems is still difficult to obtain. In addition, using a naïve approach to study these phenomena with this type of inappropriate data may yield incorrect results.

Researchers have traditionally addressed this issues by relying on data from previously

[1] http://www.youtube.com

validated surveys or questionnaires. Although a lot of progress was accomplished through these data collection methods, respondent fatigue [Porter et al., 2004] associated with survey methods and the lack of fine grained observations of the interactions (available only through other more expensive means such as observational studies) has limited the power of such studies.

Therefore, it is necessary to develop tools that can collect the desired information in a seamless and efficient way so that data can be gathered while guarantying that it is informed from established behavioral theories. This is a task that can also be characterized as within the scope of Human-Computer Interaction (HCI) and that has been, up to a point, addressed by subfields such as Human Computation, Crowdsourcing, and Games with a Purpose.

1.2 Analyzing Social Interactions

Obtaining data that is relevant for a specific problem is just the first phase that must be addressed by a system aimed at facilitating the study of online and offline behavior. Once meaningful data about social interactions is gathered, a common technique to study the relationships and interactions of people embedded in a network is Social Network Analysis (SNA).

SNA studies the relationships within nodes in a network in order to predict patterns, trace information, and discover the effects of the properties of the network itself [Wasserman and Faust, 1994]. Its units of analysis are the nodes and arcs forming the network and its focus is on the structural properties of the network such as connectivity and centrality of the nodes.

Typical applications of SNA include the prediction of links between nodes [Liben-Nowell and Kleinberg, 2007], the inference of communities [Clauset et al., 2004], and the evolution of the network across time [Kumar et al., 2010]. SNA assumes that the nodes (most typically the people represented) are interdependent, and thus, dyads (pairs of connected nodes), triads, or larger groups of nodes are of special interest.

A common approach to SNA is to assume that an observed network is simply an observation of the underlying network that occurs with a certain probability [Robins et al., 2007]. This means that in order to answer a question about the network (e.g., predicting a new tie), we need large amounts of network data to learn the appropriate parameters or previous knowledge to make educated guesses about them.

This makes doing inference in large social networks intractable because obtaining the

exact probability of a given network requires considering all possible networks that one may observe. Even if an approximation is acceptable, the estimation of parameters still require large amounts of data and, as we mentioned above, for some particular applications this data is difficult to obtain.

Alternatively, the more general problem of doing sound inference with observations relating to hidden variables has been widely addressed in the field of Artificial Intelligence (AI). In particular, probabilistic reasoning methods can be useful for SNA if we consider the network as a graphical model encoding conditional independences in the fashion of Bayesian Networks [Pearl, 1988] (BNs).

Research in BNs is extensive but it has mostly focused on developing algorithms to reason efficiently and for optimizing objective performance functions. All this while remaining agnostic to the nature of the data encoded in the network. In addition, to the best of our knowledge, there has been no particular connection to social or behavioral applications.

To address this issue, it is necessary to develop algorithms that do not only scale well with respect to large amounts of data, as is the case of most SNA applications, but that can also handle limited amounts of data, in the cases where it is scarce. Also, it is important that the results match (or at least are informed by) those that have been previously validated by behavioral sciences.

Once the data has been collected and sound inferences have been made with it, it is necessary to validate and relate the results to the offline behavior under study. This requires using statistical machinery and other high-level techniques such as visualizations, in order to further understand and explore the obtained results.

1.3 Proposed Approach

In summary, there are three main components that need to be guaranteed in order to appropriately study offline social interactions through online behavior. These are:

1. The data collected needs to be relevant to the problem at hand.
2. The inference and learning algorithms used with the gathered data need to be able to draw correct solutions when a small amount of data is provided, while still maintaining scalability when used for large data sets.
3. Tools for comparing the results from the system and other previously validated methods need to be provided in order to allow researchers to arrive at better conclusions.

In this thesis, these challenges will be addressed through the careful design of games that act as social sensors that record behavior that is both informative and relevant to the problem at hand. We propose that Social Sensing Games (SSGs) can help scientist study and analyze peer interactions in computer-mediated environments and their relation to offline behavior.

SSGs lie at the intersection of HCI and AI as they provide an application consisting of a system designed to gather data from peer interactions, and by proposing an algorithm that can draw conclusions from the data obtained through the game. We also provide a visualization tool and statistical analysis comparing the output of the SSGs with results previously validated from Social Psychology and that highlight new insights that may be drawn thanks to our framework.

As a methodology, Social Sensing Games can be classified as instances of Computational Social Science that are intended for studying people's behavior and collecting data that is relevant and meaningful for the problem for which they were designed. SSGs draw from previous research fields such as Games With a Purpose (for the design of the game interface) and from Social Network Analysis and Probabilistic Reasoning (for the algorithms developed for inference).

In order to evaluate the applicability of SSGs, we present two case studies of SSGs: the evaluation of commonsense knowledge and the identification of aggressive individuals in classroom. With the first SSG, we explore the way people recognize commonsense facts in a Turing Test scenario while taking advantage of the competitiveness between players and their shared knowledge. With the second application, the SSG helps to identify aggressive behavior and social roles in the physical world through observations done within an online social game. This second study is a collaboration with Educational Psychologists who helped design the SSG and provided the ground-truth for evaluation.

We also present additional methods to explore the output of SSGs (such as the visualizations mentioned before) and correlational analyses that highlight alternative interpretations of the data. Finally we also address some of the ethical and privacy concerns brought by SSGs.

1.4 Structure of the Thesis

In the next chapter, we present a more detailed description of the areas that inspired SSGs, including Human Computation and Social Network Analysis. We also provide some information relevant to each of the case studies, including Commonsense Knowledge and Reasoning and Psychological theories about peer aggression and bullying.

In Chapter 3, we proceed to the formal description of Social Sensing Games and provide some examples of such formulations. In Chapter 4 we introduce an inference algorithm that can be used with the output of SSGs to infer relevant information from heterogeneous social networks.

Chapter 5 explores alternative analysis for the data produced by SSGs including visualization and correlational analyses. Ethical and privacy concerns that arise from using this method are explored in Chapter 6. We conclude with a review of related work to SSGs in Chapter 7 and with the lessons learned and possible future directions for Social Sensing Games in Chapter 8.

CHAPTER 2

BACKGROUND

In this chapter, we present a brief review of the fields that provide context for Social Sensing Games (SSGs). We first introduce the field of Human Computation, relevant to SSGs as it studies the techniques and limitations of combining the computational power of humans and computers.

We also review Social Network Analysis (SNA) and in particular, two of the major methods for inference including the use of graphs and game theory. We conclude with a review of the theory behind the applications of SSGs presented in this thesis: the identification of aggressive individuals (or bullying) and the gathering and evaluation of commonsense knowledge.

2.1 Related Fields

2.1.1 Human Computation

Human Computation explores the idea of using the computational power of humans to solve problems that computers cannot solve automatically, yet. Its relevance to SSGs resides in the fact that the value of both Human Computation systems and SSGs emerges from the interaction of humans with computational systems (and both could not work without humans in the loop).

In most cases, Human Computation platforms define clear tasks that require no inference or learning after collecting the data. That is, instead of gathering knowledge for further processing by machines, Human Computation systems are aimed to solve the hard problems for the machines.

As an example, consider using a Human Computation system for identifying bullying in classrooms, an application of SSGs presented in this thesis. The task could be to show to the participants the interactions between members of the classroom and let them identify who

are the bullies, victims, or bystanders based on their observations. From the SSG point of view, we only use the game to gather information from the members of the classroom, with the objective of using the data in algorithms designed to automatically learn the appropriate labels.

Research in Human Computation usually focus on the design of the interfaces and mechanisms to gather the required human skill relevant to the computational problem to solve [Law and Ahn, 2011]. Another typical research topic focuses on which incentives to give to humans to motivate participation [Quinn and Bederson, 2011].

Examples of some of these systems use the human visual recognition capabilities (e.g., the ESP Game [von Ahn and Dabbish, 2004]), the human language understanding (e.g., Soylent [Bernstein et al., 2010]), or the human geometric capabilities and intuition (e.g., Foldit [Khatib et al., 2011]). From the point of view of incentives, there are systems that directly pay people (e.g., Crowdflower[1]), those that are intended to provide entertainment (e.g., GWAPs [von Ahn and Dabbish, 2008]), or those that use reputation (e.g., StackOverflow[2]). In contrast, SSGs are aimed to complement the work of experts (e.g., help psychologists study aggressive behavior without the need of them making the observations) in a correct and efficient way.

In the rest of this section, we will survey the field of Human Computation and its scope. We will also show how SSGs are related to existing systems but essentially address a new problem in the general area of Human Computation.

In Table 2.1, we highlight six of the main characteristics that help us distinguish between all these fields. The first one is the main focus of the systems (or what they put emphasis on). For example, Games With a Purpose are interested in participants solving a particular task whereas Social Computing platforms are interested in allowing participants to interact among themselves.

The second column is the purpose of building such systems, some are built in order to get work done from the participants (e.g., Crowdsourcing) whereas others aim to give knowledge or an experience to participants (e.g., Serious Games). Another important dimension for comparison is the type of incentives that these platforms give to participants. These incentives can be simple fun, money or reputation. Also, in the next column, we show how these systems aim to replace more expensive work that can be done by either human experts or machines. In the last two columns, we include information about how they also have a

[1]https://crowdflower.com/
[2]http://stackoverflow.com/

Table 2.1: Comparison of SSGs to other related fields in term of their objective, methods, and other characteristics. The color in the cells is aimed to highlight how each row (i.e., each field) differs from the others.

	Focuses on	Aims to	Incentives	Participants replace	Minimum Number of Participants	Participation is	Example
Games With a Purpose	Task	Get Work	Fun or Reputation	Computers	One	Voluntary	ESP Game
Crowdsourcing	Task	Get Work	Money or Reputation	Experts	Many	Voluntary	Mechanical Turk
Serious Games	Task	Give Knowledge	Fun	Experts	One	Required	IBM City-One
Gamification	Task	Get Engagement	Fun, Money, or Reputation	Experts or Computers	One	Voluntary	Foursquare
Social Computing	Interactions	Give Collaboration Space	Reputation	Experts	Many	Voluntary	Wikis
Social Sensing Games	Interactions	Get Interactions	Fun or Reputation	Experts or Computers	Many	Required	Turing Game

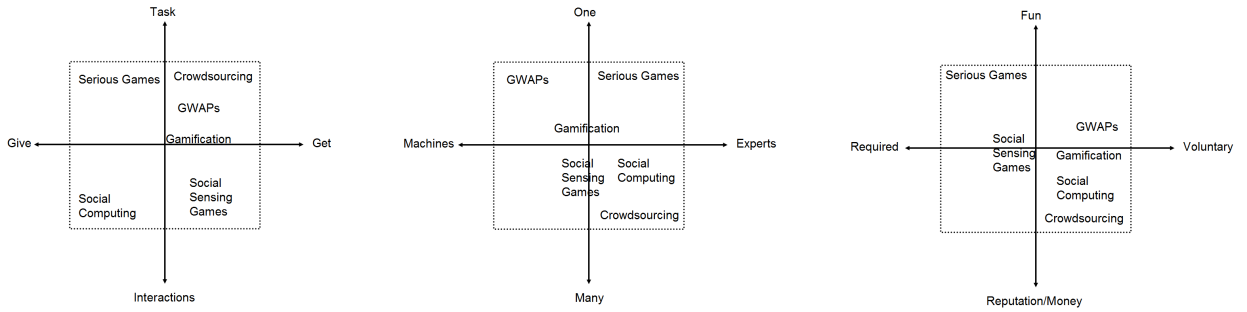


Figure 2.1: Comparison of related fields in terms of their focus and purpose (on the left), the minimum number of participants and what they are aimed to replace (on the middle), and the incentives provided and recruitment process for participants (on the right).

different minimum requirement of players and rely on different recruitment techniques that vary from open calls to requiring the participation of particular groups. Figure 2.1 shows a comparison between all these fields using the above mentioned dimensions.

History and Relation to Other Fields

The term *Human Computation* has been used as early as 1838 [Wayland, 1838], its modern use was made popular by Luis Von Ahn in [von Ahn and Dabbish, 2008] and it is now used to describe the general area that studies the power of combining humans in the computational loop to solve hard computational problems. Examples and applications of such systems include labeling images [von Ahn and Dabbish, 2004], improving security [Von Ahn et al., 2008], and solving other difficult ad-hoc challenges [Tang et al., 2011].

We further divide the field of Human Computation into the following subfields: Games with a Purpose, Crowdsourcing, Serious Gaming, Gamification, and Social Computing. The actual boundaries of these fields are not crisp as many of the applications can be conceptualized as belonging to one or several of these fields. For example, Duolingo [von Ahn, 2013] is an online website and smartphone app that teaches people different languages, it uses a gamified interface with the purpose of teaching human languages (similar to what Serious Games do), while using the work of the players to translate websites from one language to another (i.e., it can be considered a game with a purpose). Duolingo also has a social component in the form of forums and discussion boards (typical traits of social computing) and, as participation is voluntary and the final translation is aggregated from the contribution of thousands of participants, it can also be considered as a method for crowdsourcing translation.

In order to help us position Social Sensing Games in the context of Human Computation we will present a simple definition of each of the above mentioned fields to show how SSGs address a previously neglected area within Human Computation.

Games with a Purpose

The term Games with a Purpose (GWAPs) describe the type of games which are aimed to use the work of humans to help computers solve hard computational problems. The innovation in these type of games comes from the idea of taking advantage of the desire of humans for entertainment and harvesting the natural computational powers of humans to combine it with the power of computers.

The current state of AI is such that computers are very good at solving problems which humans find difficult (e.g., theorem proving, finding patterns in large amounts of data, probabilistic reasoning, etc.) while it under performs in tasks that are trivial for humans even at a very young age (e.g., commonsense reasoning, analogies, object recognition, etc.). GWAPs were inspired by the idea that humans already spend time solving many of these computational problems every day, so having a way to harvest these efforts could alleviate the computational load and provide quick and good solutions to known problems.

One of the first examples of this approach was the ESP game [von Ahn and Dabbish, 2004]. In this game, players were randomly matched with other players through an interface that did not allowed them to communicate with each other or know each other identities with one exception, they were both shown the same picture and were allowed to write one word. The players would win points if they guessed what the other person was thinking (therefore the name ESP game). Naturally, the best decision to earn points was to write something related to the image that was the only common piece of information that both players shared. This served to generate meta-data or labels for the images being displayed.

From our perspective, the main contribution of the ESP Game was to show that a clever design in a game could help (and is in fact critical) to get the appropriate data for a particular task. The design of the game had to be revised several times to stop players to abuse the games rules and achieve high scores without providing meaningful data. Some of these modifications included the use of taboo words, automatic passing of images, minimum number of labels generated before the label was actually assigned to the image, etc [Robertson et al., 2009]. Other games of a similar nature include Collabio [Bernstein et al., 2009], Verbosity [von Ahn et al., 2006], and Wikispeedia [West et al., 2009]

in which similar ideas are used to solve different problems such as social network tagging, generation of commonsense, and semantic relatedness.

Another famous example of a GWAP is Foldit [Khatib et al., 2011]. This game was designed to use human contributions to find optimal folding strategies for creating accurate protein structure models. The paper proved that such a very difficult task, usually performed by experts or exhaustive search computation, can be accomplished with game-based methods. Notice that this game could also be considered a type of gamified crowsdsourcing interface.

One of the main challenge of GWAPs is to design them in a way that still solve the computational problem at hand but that are still attractive to users as the participant voluntarily take part in the game and the only incentive is the fun (or the competition) that they will experience by playing. This is needed so that many people play the games for long periods of time (as required to generate enough amount of data). As it is expected, given the fast pace of the gaming industry, this is an effort that constantly requires innovation and updates and therefore it becomes prohibitive or meaningless to maintain these games for a long time as they suffer attrition or become increasingly expensive in resources and effort to keep current and new.

GWAPs vs. SSGs

Social Sensing Games can be considered a type of Game with a Purpose as they heavily rely on ideas of turning tasks that may cause fatigue and boredom into a fun one, with the objective of solving a task that may or may not be related to the nature of the game. The main difference is that, in contrast to GWAPs, SSGs are an attempt to learn from implicit information from the players through their explicit in-game behavior, and not to use the efforts of the players to solve a hard computational task, which make our goals and design patterns different in nature.

Crowdsourcing

Crowdsourcing refers to outsourcing hard problems to an anonymous crowd of people through an open call [Quinn and Bederson, 2011]. In that sense, games like the ESP game and FoldIt fit in the intersection of Games with a Purpose and Crowdsourcing but in general, the latter deals with a more general kind of problems.

For example, the NSF created a challenge consisting of finding 5 red weather balloons that

were somewhere in the sky of continental US [Tang et al., 2011]. The challenge was solved in a matter of hours by a team that treated the problem as that of finding information in a social network and designed a reward function that encouraged people to contact as many people as possible that could actually hold the answer.

This example represents a trend in the crowdsourcing community of implementing complex algorithms with people becoming the unit of process. It also shows how this methodology can address problems that are not necessarily computational but related to coordination.

Another example of crowdsourcing is Soylent [Bernstein et al., 2010]. In this work, the author has people respond to crowdsourced editing requests in real time and has some people deal with different parts of the problem, effectively allowing real-time editing services from an anonymous crowd.

The most known platforms for crowdsourcing are Amazon Mechanical Turk[3], CrowdFlower[4], and Kaggle[5]. They all focus in different kind of problems and provide different incentives. These are popular platforms for researchers to recruit participants to solve easy tasks for humans (but too complex for machines) such as labeling images, summarization, evaluation of query results, or to simply fill out surveys to be used for marketing and behavioral research.

The main characteristic that helps categorize an application as an example of Crowdsourcing is the reason of why to include humans in the loop. In its purest form, Games with a Purpose attempt to replace the work that ideally a computer could do by the processing power of humans, whereas crowdsourcing allow humans to do collaborative work. To exemplify this difference, consider Wikipedia[6] where the objective is to produce a set of encyclopedic knowledge (ideally from a neutral standpoint) from the collaboration of people willing to generate the content. In contrast to GWAPs, Wikipedia is not designed to fill the place of a machine but to allow collaborative writing in place of professional encyclopedia authors.

From the description above, Wikipedia is a better example of crowdsourcing for it replaces traditional human employees with an undefined, generally large group of people through an open call. There is, of course, an overlap between the two approaches described above as some systems fulfill both the definition of GWAPs and Crowdsourcing.

[3]https://www.mturk.com/
[4]https://crowdflower.com/
[5]https://www.kaggle.com/
[6]http://www.wikipedia.com

Crowdsourcing vs. SSGs

The difference between Crowdsourcing and SSGs is clear. SSGs are not interested in harvesting the computational power of people, but to learn things about them using platforms similar to those used in some crowdsourcing methods, in particular, games. Also, as was the case with GWAPs, SSGs differ from crowdsourcing in the kind of incentives that they use and the reason for including humans.

Serious Games

Serious games refer to the use of games for purposes that are not considered entertainment [Ritterfeld et al., 2009]. For example, educational games or gameful interfaces that help people save energy [Geelen et al., 2012] or to learn leadership strategies [Lopes et al., 2013]. Their main characteristic is that their purpose is not to solve a particular problem or to gather information from the user but to serve the user to accomplish or learn something.

Because of the nature of Serious Games, they do not necessarily suffer from the attrition that GWAPs do. This is because Serious Games are not intended to be sought out for their entertainment value, but because of the lesson they provide. For example, an educational game that teaches history can be used as a teaching tool by teachers who basically can force their students to play. The students then will have the benefit of the learning experience through a novel and slightly fun interface, but no one really expects people to naturally flock toward such games.

An example of a Serious Game is IBM CityOne [IBM, 2011]. It was designed to teach people how systems like smart grids and intelligent water management works. It is a sim-style game in which the players make decisions to improve a city using technology and concepts such as service reuse, cloud technologies and others.

Serious Games vs. SSGs

SSGs differ from Serious Games as their objective is not to teach or educate the players but instead they are a novel way of collecting information of the players and how they interact with one another. SSGs could include an educational component, for example teaching participants how to effectively deal with bullying, and then be considered a type of Serious Game. Instead, SSGs attempt to learn from people in order to inform behavioral theories and not to impose or change the existing behavior.

Gamification

Gamification's goal is to turn tasks that are usually considered boring or bland into fun ones [McGonigal, 2011]. It is mainly used by marketing teams to increase the engagement of customers or to incentivize certain types of actions. They can also be used to learn information from users. Nevertheless, instead of being a way of getting participants to reveal information or encourage engagement, as Gamified interfaces usually do, SSGs are aimed to learn from the users while they do what they usually would and are not aimed to change that behavior.

In its most basic incarnation, loyalty points or frequent flyers rewards may be considered a primitive type of gamification. Several websites have used different levels of gamification such as granting points, badges, or challenges to its users. The basic idea of gamification is to use of game elements or rationales in non-gaming contexts.

This field has been criticized by many experts as a simple type of exploitationware that tries to trick people into behaving in a certain way or to get them to reveal information that would otherwise be private. There is not a lot of academic work covering this type of applications although some recent papers have surfaced mostly trying to differentiate the techniques used in Serious games, Games with a Purpose, actual gaming, and Gamification [Werbach and Hunter, 2012].

Gamified interfaces do suffer from the same problem of attrition (like GWAPs) as part of their goal is to keep encouraging people to return and keep playing (or to make the players return to a website, for example). This requires the content to be updated or added often. Gamification also suffers from over simplification from people not familiar with game design concepts and so, some experts refer to many gamified interfaces as badgification or pointification as they do little else than simply award points or badges to people for doing basic things such as clicking or registering for a service.

Gamification vs. SSGs

Social Sensing Games differ from Gamification as we are not simply adding points or badges to reward certain type of behavior. In that sense, our games are more similar to other kind of research methods such as observational studies in the playground, where we are interested in observing people and how they naturally behave and not to change or direct their behavior. Also, we are not creating a gamified version of what participants already do, but creating an actual game that, through the explicit actions in the game, we can infer

implicit knowledge. Gamification on the other hand is actually interested in the explicit behavior (e.g., registering, clicking, sharing) and, usually, not on the reason why this behavior happen.

Social Computing

The term Social Computing broadly includes all systems that have humans in a social role where communication is mediated by technology [Wang et al., 2007]. Its purpose is not to perform computation or accomplish a job, but simply to facilitate collective action and social interaction online through the exchange of multimedia information and the evolution of aggregate knowledge. This definition is broad and it encompasses systems such as Wikipedia and other more general-purpose platforms such as Facebook, Twitter, Pinterest, among others.

There exist many examples of Social Computing. Systems that support of collaborative filtering [Cosley et al., 2003], open source coding [Dabbish et al., 2012], among others, are instances of the general idea behind social computing. In contrast with several of the other types of Human Computation, Social Computing relies on the interest of the participants to collaborate and engage with others with similar interest. Content is usually generated by the participants themselves and reputation systems are usually implemented to encourage creativity.

Social Computing vs. SSGs

Systems in this area are not aimed to solve computational problems but to allow participants to interact among each other. In this sense, SSGs can be considered a type of Social Computing with the added objective of collecting such interactions with the goal of learning about the participants. Nevertheless, the reason to use an SSG greatly differs to the reason to use Social Computing. Strictly speaking, SSGs objectives are not to help participants to communicate or collaborate but to study the participants' interactions.

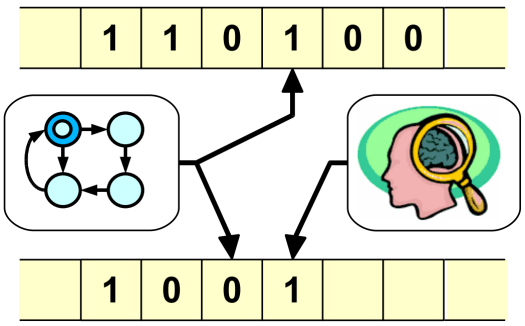


Figure 2.2: Model of a Human-Assisted Turing Machine consisting of a Turing Machine with access to a Human Oracle (included here with the permission of the authors of [Shahaf and Amir, 2007]).

2.1.2 Formalization of Human Computation: The Human-Assisted Turing Machine

To our surprise, very few attempts have been made to formalize the ideas behind human computation. One such attempt uses analogies to standard computational models such as Turing Machines while other attempts have used game theory. Part of the reason is the difficulty of formalizing concepts that include human decision making, as we humans are known to behave in rather unpredictable ways. Also, there is a strong component of design that goes into these systems which means that more than a formal proof of the adequacy of a particular game, the design elements of the platform play a role at least as important as the incentives or rational behind the design.

One of the formalizations for human computation is provided in [Shahaf and Amir, 2007]. They expand on the traditional definition of a Turing Machine (TM) by adding a human oracle which adds power to the machine. They called it a Human-Assisted Turing Machine (HTM). By having access to this oracle, the HTM can solve problems that a regular TM can't, or reduce the complexity of such tasks. For example, in the case of labeling images, which is a known hard problem, can be reduced to an algorithm that only requires linear time, given that an oracle does linear amount of work.

In their paper, the authors characterize the complexity of problems that can be solved with an interaction between humans and computers. They define a HTM as a turing machine that has access to a tape that gets filled by an oracle H (most commonly a human) that can decide certain problems, for example the labeling of an image, or general classification. See Figure 2.2 (included here with permission from the authors). The complexity of such machine is defined as a tuple $\langle \phi_H \left(M^H\right), \phi_M \left(M^H\right) \rangle$ describing the amount of work the

Table 2.2: Examples of Complexities for problem solved by HTMs. Taken from [Shahaf and Amir, 2007] with permission from the authors.

Problem	Complexity
OCR	$\langle O(1), poly(n)\rangle$
Turing Test	$\langle O(n), O(n^2)\rangle$
Classification	$\langle O(n), O(n)\rangle$

human H and the machine M need to do.

For example, assume a classification problem that is trivial to humans (e.g., face detection) where the computational task can be defined as finding the w such that $h_w(x) = 1$ if $x > w$, and 0 otherwise. The most trivial HTM to solve this problem is to show each sample to the human, this would imply a complexity of $\langle O(n), O(n)\rangle$ as all the n samples need to be observed and sent to the human which in turn labels all n samples. In contrast, we can imagine a machine that first sorts the samples according to their w score and then using binary search shows only $\log n$ samples to the human until it finds the threshold, this algorithm would take $\langle O(\log n), O(n \log n)\rangle$ for the HTM. In this example, if we consider human work more expensive than machine work, the second algorithm would be better given that the computation still takes polynomial time.

The authors also introduce different possible models for the human oracle including adding complexity to their task (i.e., not being able to decide in constant time the appropriate answer), probabilistic models, adding utility to benefit the oracle, persistence, and others. Examples of some problems analyzed with the HTM framework are shown in Table 2.2.

This formalization is appropriate when the Turing machine is solving computational problems but it is not capable of capturing the case when the objective is learning something about the oracle (in this case the human) or when there are several humans and the interest is to learn something about how they interact (which is our case for SSGs). Therefore, we will introduce a different formalization that does not rely on the concept of Turing machines.

2.2 Inference in Social Networks Analysis

As mentioned above, one of the objectives of SSGs is to gather data about a problem and analyze it. In order to solve this problem, SSGs require algorithms that can use the collected data in an efficient way.

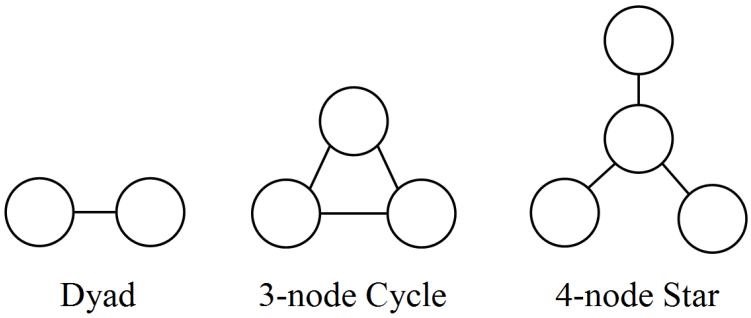


Figure 2.3: Example of common features for SNA.

The type of data gathered by SSGs, that focus specifically in peer interactions, can be considered a social network. Therefore, it is important to understand the state of the art on making inference in social networks.

We begin by giving a brief summary of one of the most common methods, namely the P* or Exponential Random Graph Models. We will also present alternative ways of studying social networks, in particular we will explain the game theoretic perspective to social network analysis.

2.2.1 P* or Exponential Random Graph Models (ERG)

Exponential Random Graph models are relevant to SSGs as its output is a social network which could be analyzed using such models. In this thesis, we also provide a simple alternative to traditional ERG models that relies on strong assumptions in order to reduce the complexity of inference and, in order to understand such benefit, it is important to briefly introduce ERG models.

For ERGs, social networks are considered graphs where the nodes and arcs of the graph represent random variables. Every instantiation of a social network is then an observation of the real social network that occurs with a certain probability. This means that, given a set of observations of the network, one is able to compute the network with the highest probability. This allows researchers to predict missing arcs that are not being observed (but should have), attributes of nodes, or the evolution of the network across time.

Because of the nature of the studies of social networks, the features used in its analysis capture the interactions between entities (or nodes). For example, common features include dyads (pair of nodes), triads (set of three nodes), stars, etc. (See Figure 2.3). These features are used to represent the network in order to capture its local structure. Depending on the

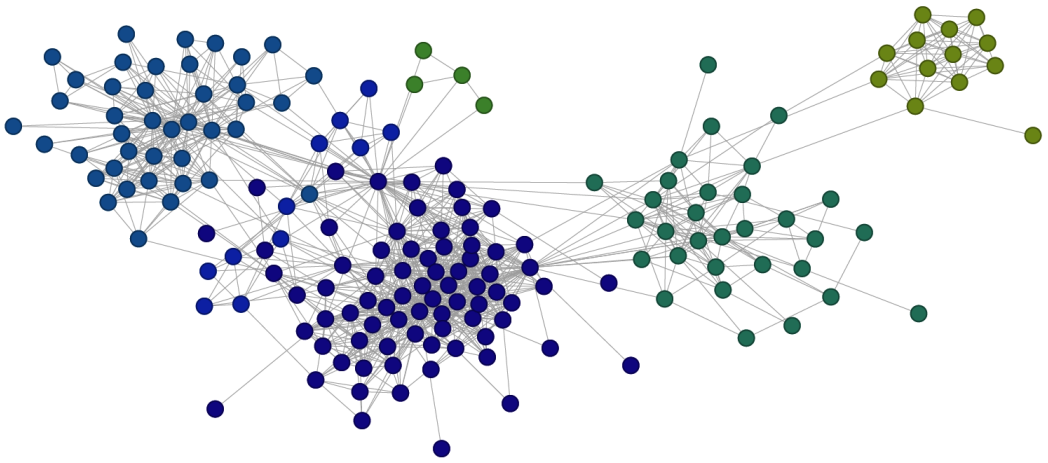


Figure 2.4: Example of a friendship network.

kind of network that is being encoded, certain features are expected to be observed more than others. For example, a typical friendship network is expected to have large cliques of people connected through a small number of nodes that serve as connections between groups (See Figure 2.4).

From the ERG perspective, a social network is defined as a set of actors (represented as nodes of a graph) and a collection of social relations (represented as edges in the graph) that specify how the actors are related to one another. The main assumption of the ERG model is that edges occur randomly [Robins et al., 2007], favoring a particular set of local structures, i.e., dyads, triangles, etc. An observed network is simply a realization of all the possible networks that can exist given the set of actors.

One of the challenges in SNA using this kind of representation is doing exact inference. In order to compute the exact probability of a particular observed network, it is necessary to compute the probability of all possible observable graphs, which is in the order of $2^{O(n^2)}$ for the case of undirected graphs with n nodes (which in most interesting networks becomes larger as there are many different type of directed and undirected arcs plus different types of nodes).

Therefore, this problem is usually addressed by the use of approximate methods such as MCMC sampling [Snijders, 2002]. These methods provide appropriate results for some cases but have proven inappropriate for predicting the evolution of networks in long periods of times [Handcock, 2003] and, because of the agnostic nature of the computation, do not exploit some of the common structure of the network that is important for certain applications [Yadati and Narayanam, 2011].

A recent alternative to ERG models (in order to avoid using sampling methods) is to use lifted inference to compute an approximation of the partition function required to compute the desired probability [Pu et al., 2012]. This is an emerging area of research which can be used in conjunction with the models used in SSGs that is currently being explored.

2.2.2 Game Theoretical Approach to SNA

Another approach to social network analysis is to focus not in the structure of the network but on the nature of the arcs being observed. One common critique to ERG models is that social networks in real life do not form at random (as is the basic assumption in ERG models) but are formed by rational agents with a specific purpose in mind [Yadati and Narayanam, 2011]. If we consider that each arc in a social network comes by two agents making a rational decision, then the formation will follow principles that can be studied through game theory. This is relevant to SSGs as the data obtained in the game is definitely a reflection of agents trying to maximize some utility (i.e., points in a game, reputation within the peer groups, etc.) and as such, could be analyzed with game theoretic models.

In [Yadati and Narayanam, 2011], the authors explain how some of the common tasks of SNA, such as node ranking, diversity among nodes, and link prediction, are not adequately done by ERG models as auxiliary information of the nodes is often required. It ignores the (possibly) strategic behavior of the nodes (or agents), and the appropriate analysis might need to include knowledge about connectivity patterns.

The main argument supporting this approach is that in most interesting applications (relating to human networks) the behavior of the agents is driven by an underlying purpose and as such, ERG models fail to capture the fact that links are formed by choice and not by chance and that there might exist incentives to forming certain links. For example, in a network formed by collaborating scientists, agents will link to each other in such a way that makes sense for their research interest (possibly maximizing their success in publishing papers). Other example of such networks might include online social networks (e.g., Facebook, LinkedIn) and telecommunication networks where the agents have a particular utility to maximize.

Naturally, some SNA tasks are more appropriate for a game theoretical approach. For example, the study of social network formation and community detection, whereas other tasks (e.g., centrality measures) are not.

As an example of how Game Theory can be used to model SNA (in particular network

formation) consider the case in which link creation has an associated value and cost. In that case, agents are expected to act strategically. The strategic network formation game is then defined as a tuple $\langle N, S_i, u_i \rangle$ where N is the set of players, S_i is the set of strategies (i.e., the set of individuals each player wants to form a link), and u_i is the utility received by the player for forming a link with a particular individual i. Following this definition, it is possible to explore the kind of networks that can be formed depending on whether links are formed under mutual consent or not.

Even though this model is simple and elegant, it depends on the assumption that the information for describing the game unambiguously is readily available and correct, which in most cases is the hardest thing to come by, and relies on the classic game theoretic assumption that agents act rationally (which in reality they don't).

In this sense, the output of SSGs could serve as input data for game-theoretic models of SNA as the data of a real network is used and data about the possible utilities is gathered. Other relevant work in game theoretic models for SNA include [Jackson and Wolinsky, 1996, Buskens and Van de Rijt, 2008].

2.3 About the Applications of Social Sensing Games

2.3.1 Identification of Aggressive Individuals in Classrooms

Psychological Theories of Aggression and Bullying

The influence of peer groups in individual behavior has been vastly studied in the field of psychology. Example of such influence includes smoking [Ennett et al., 2008] and binge eating [Crandall, 1988]. Because of the application presented in this thesis, of particular interest to SSGs is the influence that peers have on aggression and bullying [Espelage and Holt, 2001].

Some theories of aggression explain the differences between bullies and victims with the help of internal characteristics (such as executive functions [Monks et al., 2005]) whereas others use social constructs [Espelage et al., 2004]. The former suggests that there are differences between the executive functions of bullies and victims as well as differences in terms of how much attention they pay, how interactive they are, and how many prosocial interactions they pursue, suggesting differences between aggressors and victims.

Theories that address bullying (and cyberbullying) from the social standpoint are more

relevant to SSGs as the application presented in this thesis relies on social interaction for the identification of bullying. One of these theories is called the social-ecological theory [Bronfenbrenner, 1977, Espelage and Horne, 2008] that identifies associated risks and protective factors across all context.

According to the social-ecological theory, children and adolescent behavior is shaped by a range of nested contextual systems, including family, peers, and school environments. A child's direct contact with family, peers, and schools comprise the *microsystem*. Parent-teacher meetings are an example of a *mesosystem*. The *exosystem* is the social context with which the child does not have direct contact, but which affects her indirectly through the microsystem. The *macrosystem* comprises influences from a child's larger environment such as cultural values, customs, and laws [Littlefield-Cook et al., 2005]. Finally, the dimension of time is called the *chronosystem*. SSGs take some inspiration from this theory as they help scientists gather information about the school environment of the children (the microsystem) and infer the roles of each of the participant in the classroom social network.

Another theory is the Social Information Processing Deficit Model [Dodge et al., 1986, Dodge and Coie, 1987], which proposes that aggression is largely due to impairment in social problem solving. This complex model has been supported across a multitude of studies and concludes that aggressive children tend to show encoding problems such as hostile attribution error, deficits at the level of representation (e.g., a poor understanding of others mental states), and a limited solutions to social problems. More specifically, the authors found that children who behaved aggressively were more likely to attribute hostility to ambiguous situations and thereby have deficits in interpreting social information.

Supporters of this theory argue that a child's behavior is directly related to his mental processing of the situation and competent social information processing results in adaptive and competent prosocial behavior. SSGs can help us provide evidence for this theory as the tasks that are created by the SSGs for the participants to solve can help us observe their problem solving strategies and outcomes, as well as their interactions with peers.

From a social learning perspective [Bandura, 1986], the external environment is considered to contribute to the acquisition and maintenance of aggression and other risky behaviors. The development of aggression is believed to be the consequence of exposure to socially deviant role models and inappropriate reinforcement of risky behaviors. For example, a child who bullies other children and receives reinforcement from other students who laugh, join in, or generally offer support to the bully will be likely to continue his behavior. What is not clear is whether this reinforcement serves the same type of function in cyberspace. For

example, it is plausible that engaging in cyberbullying within social network sites might also provide reinforcement for such behaviors by others liking a post etc. SSGs can help answer such questions as they are capable of observing when aggression occurs and the reaction of peers to such aggression (in terms of whether it is accepted, encouraged, or rejected).

Another relevant theory of aggression is Resource control theory (RCT) [Hawley, 2003]. By resources it is meant the material, social, and informational things that are generally seen as desirable by children. Hawley argued that it is likely that the social skills deficit model might not account for the adaptive function of bullying and the utility of being a bully among youth. RCT argues that individual youth have the capacity to employ a wide range of strategies when interacting with their peers to obtain limited resources. More specifically, they use coercive and pro-social strategies to obtain resources, and in doing so foster social dominance. In [Hawley et al., 2007], these individuals are called bi-strategic controllers within school classrooms for they use prosocial (friendly) and coercive (threatening) strategies in order to obtain resources. Again, SSGs can help us test this theory as we can simulate the presence of limited resources and monitor the strategies used by the participants to obtain and maintain such resources.

Online Disinhibition Effect

An important factor to consider when studying bullying and aggressive behavior in online or virtual spaces (as is the case in SSGs) is the so called Online Disinhibition Effect. In some ways, the cyberspace provides individuals increased access to their social network and, unlike face-to-face interactions, individuals can engage with a large number of individuals at once at any given time. Nevertheless, even though adolescents have an increased ability to interact with their peers, social interactions through technology lack the immediate and feedback that are inherent in face-to-face interactions.

In [Suler, 2004], the author described a phenomenon called the online disinhibition effect, which refers to greatly diminished internal censorship when communicating in cyberspace. He claims that "*people say and do things in cyberspace they would not ordinarily do in the face-to-face world. They loosen up, feel less restrained and express themselves more openly*". This effect can be either benign (encouraging appropriate self-disclosure) or toxic (encouraging mean or cruel attacks on others).

Suler proposed that six factors inherent in the technology contribute to this effect: dissociative anonymity (which allows one to mentally separate online activity from real life by

concealing ones identity), invisibility (inability to see or be seen by those with whom one is communicating) , asynchronicity (allowing one to avoid knowing the receiver's immediate reaction to a communication), solipsistic introjections (incorporating an imagined receivers personality into ones own psyche), dissociative imagination (the belief that the personas one creates in cyber-environments remain in an online world, limiting responsibility for real-world consequences), and minimization of authority (because the usual markers of status are absent in cyberspace). This tendency to exhibit a more narcissistic, aggressive, and uncivil persona in the digital world is also described in [Aboujaoude, 2012], where they propose that a more dangerous personality exists parallel to our non-digital selves.

Therefore, in virtual environments, people can choose to interact with others anonymously, and avoid the repercussions that might accompany the bad behavior if they were identifiable. This might encourage people to say or do things online that they are unlikely to do in their face-to-face interactions and to limit their sense of responsibility for these actions [Blumenfeld, 2005].

Researchers found that in investigations of cyberbullying, perpetrators reduced their sense of responsibility for the abusive nature of their online messages using rationalizations centered around offering the targets of their abuse needed and useful information. For example, when asked why they sent abusive messages to others online, perpetrators responded "*I was only telling the truth. She is ugly, and I felt she had to know it!*" Thus, online disinhibition theories offer the field a potential explanation for involvement in cyberbullying.

The SSG created for bullying identification take this effect into consideration by restricting its use to the context of the physical classroom of the participants and disclosing their identity to all players. This helps the anonymity that causes online disinhibition and the sense of invisibility to disappear, allowing for a greater connection between the online persona and the behavior offline. This prevents the online disinhibition effect and allows us to obtain meaningful data.

2.3.2 Commonsense Knowledge and Reasoning

One of the long standing goals of Artificial Intelligence is to create systems or agents that have human-level intelligence. This includes the capability of handling every day actions and having the knowledge (or the capability of inferring) common things that humans can do. This is usually refer to as commonsense and has proven to be a very difficult topic to handle from the formal point of view.

One of the main challenges of Commonsense is the gathering and evaluation of knowledge. In this thesis, we present a game aimed to collect the shared knowledge of people to evaluate knowledge as commonsense.

The study of commonsense can be divided into two major approaches that may be called commonsense knowledge (CSK), what everybody knows, or Commonsense Reasoning (CSR), "the human ability to use commonsense knowledge" [McCarthy, 1990]. From the purest CSK perspective, the problem of commonsense can be reduced to having a complete and useful commonsense knowledge base that includes factoids and other type of knowledge that allows a system to reason with commonsense. CSK can be though of as the set of commonly shared knowledge that enables much of the discourse among people in describing, predicting, or explaining everyday events. It preludes formal education and its acquisition is related to the perception of causality, the persistence of objects and properties over time [Elio, 2002].

Commonsense Knowledge

The main challenges of CSK are 1) to collect large amounts of knowledge (i.e., true useful factoids), 2) the evaluation and verification of such data, and 3) the efficient use (or inference) of such data. Several researchers have focused on the recollection of such useful data, using methods like crowdsourcing [Singh et al., 2002], games with a purpose [von Ahn et al., 2006], or hand-crafting data by knowledge engineers [Lenat et al., 1990]. Nevertheless, very few studies have focused on the verification of the data as useful or truly representative of commonsense. In this thesis, we propose a SSG built with the main purpose of evaluation of previously gathered commonsense knowledge.

Examples of methods for generating the large amounts of commonsense knowledge required for many of its applications include CyC which uses knowledge engineers to code facts in a predetermine language, OpenMind/ConceptNet [Singh et al., 2002, Liu and Singh, 2004], which uses volunteer contributors to enter facts about the world in natural language, and automatic methods such as [Matuszek et al., 2005] or NELL (Never Ending Language Learning) [Carlson et al., 2010] that extract knowledge from the web. Some of the application of commonsense knowledge include improving search engines [Liu et al., 2002], data alignment [Reed et al., 2002], and other NLP applications [Varga et al., 2010].

Commonsense Reasoning

From the CSR perspective, the problem is boiled down to finding the right logic system to reason in a commonsensical way. The field of Commonsense Reasoning attempts to write down (axiomatize) what everybody knows about some domain and apply formal proof procedures. It basically approaches human-level AI through the use of mathematical logic as a formalization of commonsense knowledge in such a way that commonsense problems can be solved by logical reasoning [McCarthy, 1968, McCarthy, 1990]. Currently, the SSG created for collecting and evaluating CSK does not include any reasoning mechanism but this section has been included for completeness.

Some of the aspects that have been studied from the CSR perspective is nonmonotonicity, an essential part of everyday commonsense [Brown, 1986]. Nonmonotonicity can be explained as the capability of retracting statements or beliefs previously made, after new information has arrived. The canonical example is that of a knowledgebase consisting of an axiom stating that birds fly. If this knowledgebase is told that tweety is a bird, it can conclude that tweety flies, but after knowing that tweety is abnormal in some sense (e.g., dead, had its wings cut out, or it is a pinguin), then it can retract the previous conclusion and state that tweety cannot fly.

There exists several approaches to nonmonotonicity, the most popular might be default logic or reasoning [Lifschitz, 1995]. This approach assumes the existence of default propositions that are assumed to be true in the absence of information to the contrary. The three basic theories of nonmonotonic logic are further explained in [McCarthy, 1990, McDermott and Doyle, 1980, Reiter, 1980].

In reality, most CSK and CSR systems fall within both CSK and CSR. Systems that include large amounts of commonsense knowledge usually include some mechanisms to soundly reason with the data, and although there are many papers dealing only with the logic of commonsense, the pre existence of a dataset that fits the logic is always assumed.

CHAPTER 3

SOCIAL SENSING THROUGH GAMES

In contrast to computer games that have been previously used to solve computationally hard problems (e.g., image labeling [von Ahn and Dabbish, 2004]). We propose that games can be used to allow participants to interact with one another in order to reveal important information about their role within their social group. In other words, games can be used as social sensors whose output is a simplified representation of participants' interactions, and that can be used for learning and reasoning about their social roles and behaviors.

3.1 General Notions

When designing a sensor, it is useful to first describe its input. In our case, we are interested in learning about individuals embedded in a social network in the real-world, henceforth referred to as *RSN*. We assume that individuals in the *RSN* (usually represented as nodes in social networks) have a set of attributes attached to them (such as age, interests, etc.) and that are mainly connected in two different ways: through relationships and through interactions.

- Relationships: We understand relationships as connections that exist between individuals that are (mostly) permanent. These can be represented by categorical variables because they stand for concepts like friendship, family, attendance to the same school, etc.

  These connections tell us something about people in the long term but might be insufficient to infer behaviors in the *RSN*. A classical task related to relationships would be to predict them in terms of the attributes of the individuals in the *RSN*. From a social network perspective, these connections can be represented as undirected unweighted arcs.

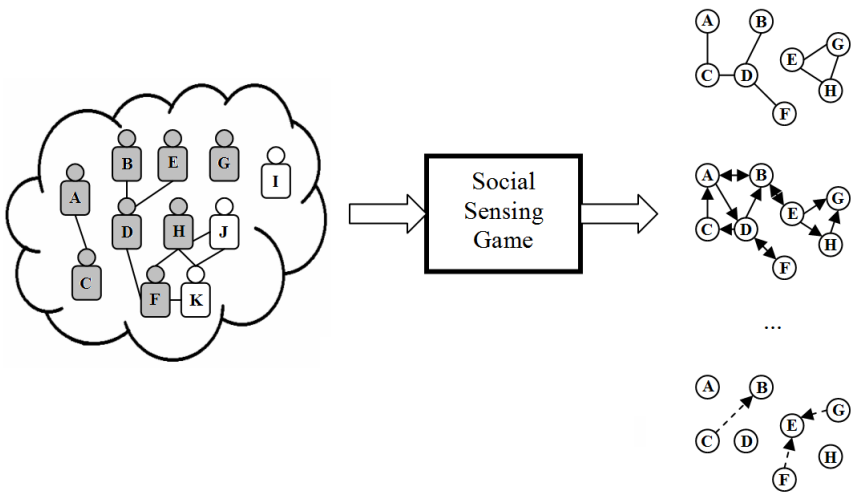


Figure 3.1: Diagram of a Social Sensing Game, its input is a social network (from the real world) and its output a representation of such network as a heterogeneous social network.

- Interactions: We define interactions as connections between individuals that vary greatly depending on when the observation occurs. These can be represented as numerical variables and they stand for concepts such as sent and received text messages, number of *Likes* in posts, frequency of phone calls, etc. Classic tasks include the inference of the nature or content of the interaction, and the number of interactions. In social networks, these connections are often represented as directed and weighted arcs.

Notice that in the *RSN*, we expect relationships and interactions to be somehow correlated but not necessarily determined by one another. For example, we would expect friends to call each other often, but we can also expect to observe phone calls between strangers and acquaintances. Also, we cannot expect all individuals in the *RSN* to take part in our data collection, i.e., we might not observe all the relationships and interactions that exist.

Ideally, the SSG takes a snapshot of the *RSN*, by observing all relationships, interactions, and attributes in a particular moment in time (see Figure 3.1). Notice that if the same *RSN* plays the game in different moments, the relationships and attributes should mostly remain the same, but the interactions should be completely different.

The output of the SSG should be a network representing all the observed information. This means that the output should be equivalent to a heterogeneous social network *HSN* inferred from the observations obtained through the game.

An *HSN* is a network that encodes many types of connections between individuals. Formally, an *HSN* can be considered as a set of graphs $G_i = \langle V, E_i \rangle$ where V is a set of nodes (the same for all G_is), each node representing an individual, and $E_i \subseteq V \times V$ representing a type of connection between two nodes.

Notice that the *HSN* is only a snapshot of the *RSN* and may not contain all the information encoded in the latter one. The *HSN* is only an approximation of the *RSN* and its quality will depend on the design and on the observations made by the sensor (in our case, the SSG). Also, notice that some information of the *RSN* might be misread, making the *HSN* correct only with a certain probability.

3.2 Definition of Social Sensor

The SSG is a game designed to collect information about the real world and to produce an approximate representation of it. Its input is information from a *RSN* and its output is an *HSN*.

Therefore, the SSG can be considered a social sensor that takes measurements of the real world. We can represent this as a function h that goes from the space of *RSN* to the space of *HSN*.

$$h : X \rightarrow Y \tag{3.1}$$

where X is a *RSN* and Y is a *HSN*.

Notice that h is a many-to-one relationship, i.e., several X can be mapped to the same Y. Also, because X is an *RSN*, we must address how h obtains information from the many individuals that form the *RSN*, their relationships, interactions, and attributes.

The game part of the SSG is designed to serve as an interface where ideally all individuals of the *RSN* take part and that is capable of capturing the attributes of the individuals, as well as their relationships and interactions. In practice, it is possible that not all the individuals of the *RSN* play the game and that the capabilities of the game to collect information are limited by its design and the medium in which it is embedded, i.e., the SSG may not capture all attributes, relationships, and interactions.

This means that h must be refined to a function that takes the information of n individuals and their interactions, as well as any self-reported or measured relationships to output an *HSN* with the appropriate information. Formally, we have

$$\begin{aligned} o &: X \to A \times R \times I \\ h &: o(X) \to Y \end{aligned} \tag{3.2}$$

where o is a function that takes the *RSN* and produces a set of observable (or measurable) attributes A, relationships R, and interactions I. h now takes the sets A, R, and I to produce Y which is an $HSN = \{G_1, G_2, ..., G_m\}$ where $G_i = \langle V, E_i \rangle$ with $V = \{v_1, v_2, ..., v_n\}$ for n individuals in the input *RSN*, and where $E_i = \{(v_j, v_k)\}$ corresponds to one of the m types of connections in the *RSN*.

The function o is determined mostly by the environment in which the social sensor exists and only slightly by the design of the sensor, whereas h is completely defined by it.

Apart from providing a convenient interface for reading information from the *RSN*, the game interface creates incentives for participants to share information while playing the game and thus revealing their attributes, relationships, and interactions.

In the next section, we formally introduce how h is implemented as a game. This will be done in terms of a task in which two or more individuals compete or collaborate to solve a particular problem while remaining agnostic as to what the specific task is (it will be defined by the specific problem the designer is trying to address) and to what *winning* the game means.

3.3 Definition of Social Game

Formally, a game $\mathcal{G}$ is defined as a tuple formed by a set of players $\mathcal{P}$, where $|\mathcal{P}| = n > 2$, a non-empty set of both *asocial* or *social* actions $\mathcal{A}$, a task $\mathcal{T}$, which is a sequence of actions $\alpha \in \mathcal{A}$ (the task of the game is said to be completed once the sequence of actions is observed), and a reward function $\mathcal{R}$ that takes a sequence of actions and returns a real number,. Each player may have his own reward function, and the reward may depend on the actions of all players, i.e., $\mathcal{R}_k : \mathcal{A}_1 \times \mathcal{A}_2, \times ... \times \mathcal{A}_n \to \mathbb{R}$, where $\mathcal{R}_k$ is the reward of the kth player defined by $\mathcal{A}_i$, the actions of the ith player. The attributes and relationships of the individuals may be collected by directly asking the players, obtained from the medium where the game is

embedded, or inferred through the actions of the players. In summary,

$$\mathcal{G} = \langle \mathcal{P}, \mathcal{A}, \mathcal{T}, \mathcal{R}_1, ..., \mathcal{R}_n \rangle \tag{3.3}$$

By *social* actions, it is meant actions that are directed towards other players (i.e., interactions) in contrast to *asocial* ones which represents actions that are performed by a player but directed to no one in particular (e.g., submitting an answer, bidding for a particular resource, etc.). This definition is meant to help designers focus on the relevant design elements of the SSG. Its expressiveness is equal to an N-player stochastic games but instead of highlighting the states of the game, we focus on the actions (the states of the game can be generated automatically by enumerating all possible sequences of actions).

The social game is useful in the same way that deception is useful in psychological studies. By carefully choosing the task to solve, we can focus the attention of the players on specific parts of the game while still providing useful and relevant information (players might not answer truthfully and objectively to direct questions).

3.4 Definition of Social Sensing Game

Social Sensing Games are then defined as the combination of both a Social Sensor (as described by the function 3.2) and a Social Game (described by the tuple 3.3). In practice, we can consider the game $\mathcal{G}$ as an implementation of h that, through the set of actions, task, and rewards, measures the attributes, relationships, and interactions embedded in the social network.

Because of the previous definition, notice that SSGs cannot address every problem with a single design. The particular characteristics of the game and the environment (i.e., A, R, and I) need to be defined in a case by case basis. SSGs allow us to unintrusively observe the real world and to tailor our data collection tools to address specific social-behavioral problems. The task $\mathcal{T}$ and the rewards $\mathcal{R}_i$ can be designed in ways that motivate certain types of interactions and improve the likelihood that observations are relevant to the problem at hand.

Example of a SSG Definition

Our definition of SSGs is such that many current social media instances can be considered a type of social sensor games, which is consistent with previous work that suggests that social media can be analyzed as games [Cirucci, 2013]. As an example, we show how Facebook can be considered a SSG that can be used to address certain types of problems.

The output of the "Facebook SSG" would be a graph Y that can be used to study problems such as: inference of attributes (such as interests), prediction of relationships (such as suggestions of friends), suggestions of interactions (reminders to poke or to tag someone), etc.

Notice that given the information that Facebook collects, it is not appropriate to infer the role that people have in the *RSN* due to the fact that people curate their profile to project an image of themselves that does not necessarily match their image in the *RSN* [Hogan, 2010]. This could be changed if the task or the rewards are modified in order to encourage truthful and complete information (e.g., using incentives like those in LinkedIn, which promote accurate depiction of the members' professional activities).

Example: Facebook as a Social Sensing Game

$$
\begin{aligned}
X &= \text{The real-world social network of Facebook users} \\
o(X) &= A \times R \times I, \text{ where} \\
A &= \{\texttt{demographics}, \texttt{interests}, \texttt{hometown}, ...\} \\
R &= \{\texttt{friends}, \texttt{acquaintances}, \texttt{married}, ...\} \\
I &= \{\texttt{talking}, \texttt{sharing}, ...\} \\
Y &= h(o(X)) = \text{Graph with all Facebook's information} \\
\mathcal{G}_{Facebook} &= \langle \mathcal{P}, \mathcal{A}, \mathcal{T}, \mathcal{R}_1, ..., \mathcal{R}_n \rangle \\
\mathcal{P} &= \text{Users of Facebook} \\
\mathcal{A} &= \{\texttt{like}, \texttt{share}, \texttt{post}, \texttt{tag}, \texttt{friend}, \texttt{unfriend}, ...\} \\
\mathcal{T} &= \text{Connect with people} \\
\mathcal{R} &= \text{Depends on the player, for some it might be the number of } \texttt{likes} \text{ received,} \\
&\quad \text{or the number of pictures posted, or the number of friends, etc.}
\end{aligned}
$$

In a sense, if we consider Facebook a Social Sensing Game, we could see that it basically attempts to create a one to one match between the properties of the real world network and its internal representation, i.e., map relationships to friendships, attributes to the profile and

interactions to Facebook's actions such as liking and commenting.

3.4.1 Conclusion

Notice that the design of SSGs are useful only for certain applications. Our goal is to highlight the characteristics that are needed to create SSGs when addressing specific problems. We want to show what games can do (when used as social sensors), and to pinpoint what designers must think about from the point of view of the mechanisms of the sensor. Also, we believe that SSGs have the advantage of not being too "general-purpose" as generic social media and can easily be adapted to solve different kind of problems.

Next we will show two studies in which SSGs are used to evaluate commonsense knowledge and to identify bullying in classrooms, respectively.

3.5 Application: Evaluating Commonsense Knowledge

3.5.1 Motivation

Collecting commonsense knowledge (CSK) [McCarthy, 1968] from freely available text can reduce the cost and effort of creating large knowledge bases [Banko and Etzioni, 2007]. Collecting commonsense knowledge is difficult because it is dynamic and dependent on context. This makes it impossible to generate it randomly and to verify it automatically, which implies that humans are needed to either collect the commonsense knowledge or to verify it.

For the acquired knowledge to be useful, it is important to ensure that it is correct, and that it carries information about its relevance and about the context in which it can be considered commonsense. To deal with the problem described above, we designed an SSG that captures the shared knowledge of people and, using this information, classifies the knowledge as commonsense (i.e., shared knowledge), domain specific (i.e., shared by only few), or nonsense [Mancilla-Caceres and Amir, 2011].

The idea behind this design is that commonsense can be considered as a shared model of the world, i.e., what two or more people know (or assume) to be true. This means that we do not care about the content of the knowledge itself, but about whether or not the opinion about that knowledge is shared by many. This SSG contributes to the field of CSK by providing a new definition of commonsense and by giving guidance about what knowledge

to use, when, and how.

This game leverages the structure of an underlying social network (in this case in particular Facebook) to find knowledge that is shared between the players and labels knowledge as commonsense if it is shared by many of the users, domain-specific if it is shared by few, and nonsense if it is meaningless or grammatically incorrect.

One common challenge when needing humans to process or generate knowledge is that encouragement (or rewards) need to be carefully designed so that even knowing that participants will only care about maximizing them, they still produce the desired outcome. Previous attempts have been only marginally successful at this because they either provided no encouragement at all, they became expensive as they paid people to collaborate, or they designed rewards allow people to maximize them regardless of whether they are producing the right data or not.

In this section, we introduce a SSG called *The Turing Game* that takes text extracted automatically from the Web and uses input from players to verify it as commonsense by comparing what knowledge they share.

As previously mentioned, the main difficulty of this approach lies in the fact that the game needs supply the correct information to solve the problem at hand. In this sense, the design of the game is much like designing an algorithm [von Ahn and Dabbish, 2008]. We solve this by focusing our design on the correctness of the data and the features that increase the replay value of the game (in order to collect enough data).

3.5.2 General Notions

We will distinguish between two types of knowledge: *commonsense knowledge* and *domain-specific knowledge*. Domain-specific knowledge includes that which is relevant to a specific group about a specific subject, the carrier of such knowledge is regarded as an expert (but notice that within that specific group, domain-specific knowledge may be considered commonsense). Commonsense knowledge is what is known by everybody everywhere, regardless of group memberships.

With the help of the players, our game classifies the knowledge as commonsense (most players know if the fact expressed by the sentence is true or false), domain-specific (only a restricted amount of players know about the fact), or meaningless (nonsense produced probably by an error of the parser used to extract the sentence). It also reports if more information about a given fact is needed in order to classify it correctly.

Correctness of the data is ensured through the design of several stages in the game, and through restricting communication among players. The reward for playing the game comes from the intellectual stimulation and sense of competition obtained through verifying the sentences and competing for high scores.

Because not all commonsense is equally common, it is not appropriate to classify all facts in a categorical fashion as either commonsense or not. Therefore, we create a continuous scale based on a Binomial Hypothesis Test that reports a degree of confidence about the decision of identifying a sentence as commonsense. This scale also allows us to clearly identify which knowledge needs revision.

The main contribution of this SSG is that it presents a method for data collection that guarantees that the data is correct by providing context that emerges from the social interactions. For this game, the source of the original data is *Simple Wikipedia*[1]. This is because some of its policies regarding the content of the articles are appropriate for commonsense extraction. For example, original research (which is clearly not commonsense) is not allowed to be part of any article.

3.5.3 Design of the Turing Game as an SSG

Before formally describing the desing of the Turing Game, we will discuss some of the circumstances and rationales of this application. The initial input information for the game comes from an off-the-shelf parser [SigArt, 2009] that extracts a sentence from an article in Simple Wikipedia and, together with the action of the user, produces an update to the knowledge base as the output. See Figure 3.2.

With regard to the task, the simplest implementation of the game would have a single user classifying sentences as either commonsense or not. This is not enough because it would be impossible to evaluate the answers of the player as correct or incorrect and it would disregard the notion that commonsense emerges from shared knowledge between two or more people.

Also, having several players evaluating the same sentence and accepting the input only if they agree amongst one another is not appropriate because it would be easy for a group of players to act in collusion and agree on entering the same answer, regardless of the question.

The basic problem is that human players can always agree on a fixed strategy, and yes/no questions are not enough to correctly classify knowledge. To solve this, we added a non-human player to the set of players and classify the input text in four different categories:

[1] http://simple.wikipedia.org/

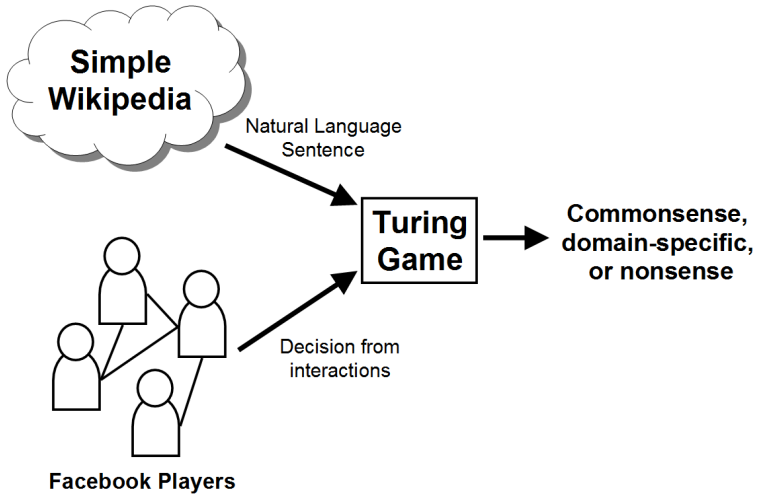


Figure 3.2: Conceptual design of the *Turing Game*. The game receives as input a sentence from Simple Wikipedia and a decision from the interactions of the players from Facebook. Its output is the evaluated knowledge contained in the sentence.

Nonsense, Unknown, True, False.

These three player game specifies as goal to distinguish between another human and a machine. While it is expected that both humans will give the same answer on a sentence, the machine-player can only guess its answer. This design of the game solves the problem outlined before: The two humans can no longer agree on any strategy because the identity of the players is unknown. Also, if the answer of one player does not follow commonsense, the other human might erroneously identify the player as a machine, which results in a penalization on the player's score.

Also, to completely define the task in a way that we get the appropriate results, and because commonsense depends on the context, it is necessary to specify the context explicitly. In [McCarthy, 1989], the author proposes a formula *Holds(p,c)* to assert that the proposition p holds in context c. Using this idea, the appropriate task for the player is to answer a question based on that formula. In our case, the context is handled by the name of the Simple Wikipedia article used as source for the sentence. The context can be used by the player to answer correctly, while addressing the problem of uninstantiated sentences that may be produced by the parser (e.g., it is impossible to know whether the sentence "*It is red*" is true or not without knowing what "*It*" refers to).

SSG Definition

According to the previous discussion, in the case of the *Turing Game*, the input X is the social network in which the participants are embedded (in this case, Facebook) and so, the attributes, relationships, and interactions are the ones found in Facebook (e.g., hometown, gender, frienships, likes, posts, etc.). Because our goal is to obtain shared knowledge, the output of the game should be an *HSN* where the nodes are connected according to what knowledge they share. For this particular case, we are not really interested in the network but in the knowledge that connects people.

The actions of the game, as discussed above, must be labeling each piece of knowledge as true, false, unknown, or nonsense. The task is to distinguish the other human player from the machine player, and the reward for each player depend on the success on the task. In summary,

The *Turing Game* as a Social Sensing Game

$$
\begin{aligned}
X &= \text{The player's Facebook social network} \\
o(X) &= A \times R \times I, \text{ where} \\
A &= \{\texttt{Facebook profile}\} \\
R &= \{\texttt{friends, friends of friends,...}\} \\
I &= \{\texttt{Like}, \texttt{share}, ...\} \\
Y &= h(o(X)) = \text{Shared knowledge} \\
\mathcal{G}_{TuringGame} &= \langle \mathcal{P}, \mathcal{A}, \mathcal{T}, \mathcal{R}_1, ..., \mathcal{R}_n \rangle \\
\mathcal{P} &= \text{Users of Facebook} \\
\mathcal{A} &= \{\texttt{True}, \texttt{False}, \texttt{Unknown}, \texttt{Nonsense}, \texttt{Identify Machine Player}\} \\
\mathcal{T} &= \text{Distinguish between human and machine player} \\
\mathcal{R} &= \text{Points are earned for correctly identifying the human player and lost} \\
&\quad \text{for any mistake or for being identified as a machine player}
\end{aligned}
$$

Figure 3.3 shows screenshots of the implementation of the game in all its stages. The game flows as follows:

- In the first stage, the player chooses a topic (which matches the title of the Wikipedia article from which a sentence is to be retrieved).

Figure 3.3: Screenshots of all the stages of the *Turing Game*. Each stage is shown one below the other.

- A sentence is randomly selected from the article. The system chooses either a new sentence or a sentence that has been verified before. This balances the coverage and reliability of the data by increasing the times a sentence has been verified. Then, the player indicates whether the fact expressed by the sentence is true in the context of the article using the four options previously described.

- In the second stage, the player sees the answer of the other two players and identifies which of the two is the machine that is answering randomly. In the case of a single player playing the game, the other answer comes from recorded games. If it is impossible to distinguish between the two players, there is an option to pass and avoid making a decision. If the player identifies the human as the machine, points are deducted; otherwise, points are awarded.

- After this, the player gets the opportunity to play again.

3.5.4 Analysis of the Collected Data

Our goal is to label each sentence as commonsense, domain-specific, or nonsense. Notice that a majority vote is not enough to generate the label because we have more confidence if a sentence was evaluated by a large amount of players rather than by a few. Thus, we create a scale of commonsense that describes how common a specific fact is. The scale needs to be proportional to the ratio of people who know the given fact, and also contain information about the confidence of such ratio. We first define four quantities, $t_{\texttt{count}}$, $f_{\texttt{count}}$, $u_{\texttt{count}}$, $n_{\texttt{count}}$, that hold the number of times a sentence s has been classified as true, false, unknown, and nonsense, respectively.

Definition 1 Let $P_\sigma(s)$ be the ratio of people that have answered true, false, and unknown over the total number of instances the sentence s has been verified. That is,

$$\begin{aligned} P_\sigma &= \frac{t_{\texttt{count}}(s)}{m(s)} \text{ , where} \\ m(s) &= t_{\texttt{count}}(s) + f_{\texttt{count}}(s) + u_{\texttt{count}}(s) + n_{\texttt{count}}(s) \end{aligned} \tag{3.4}$$

Under the assumption of independence, each instance of the game can be considered a Bernoulli trial, $P_\sigma(s)$ is an estimator of the real proportion of people that understand the sentence. Our null hypothesis is that the ratio of people classifying the sentence as nonsense

Sentence Id	Sentence	Article
1	*"People are known acting in comedies are comedians"*	Comedy
2	*"Computers can use many bits"*	Computer
3	*"For example, some languages (e.g. Chinese, Indonesian)"*	Verb

Sentence Id	Times played	$P_\sigma(s)$	p-value	Degree assigned by scale	Meaningful
1	1	1	1	0.5	Unknown
2	6	1	0.0313	0.9844	**Yes**
3	6	0.1667	0.0313	0.0156	**No**

Table 3.1: Examples of sentences played on the game. The sentence Id is used for reference in this chapter. $P_\sigma(s)$ is the proportion of people that didn't answer *nonsense*. The p-value, $p_n(s)$, corresponds to the one obtained by the Binomial Hypothesis Test. The **degree assigned by our scale**, $\pi_s(s)$, represents the confidence that we have when classifying the sentence as meaningful. The last column is the decision made regarding the meaning of the sentence with a significance of 0.1

should be 0.5. If we fail to reject the null hypothesis, we conclude that we don't have enough information to identify the sentence as meaningful or nonsense.

Definition 2 Let $e_n(s)$ be the effect size, i.e., the difference between the actual and expected number of times the sentences have been marked as nonsense.

$$e_n(s) = \left| n_{\texttt{count}}(s) - \frac{m(s)}{2} \right| \tag{3.5}$$

Definition 3 Let $p_n(s)$ be the p-value of the Binomial Hypothesis Test, the probability of observing a difference in the value of a random variable of at least the size of the effect size $e_n(s)$.

$$p_n(s) = P\left(X < \frac{m(s)}{2} - e_n(s)\right) + P\left(X > \frac{m(s)}{2} + e_n(s)\right) \tag{3.6}$$

The p-value $p_n(s)$ is the probability of observing the current counters given the null hypothesis. The lower its value, the more confident we are about their values.

Table 3.1 shows some sentences with their corresponding $P_\sigma(s)$ and $p_n(s)$. Notice that $P_\sigma(s)$ and the p-value cannot distinguish amongst all sentences because one only considers the ratio of people that agree on the sentence, whereas the other only considers the amount of people that has evaluated the sentence. With this in mind we define $\pi_s(s)$, which allows us to easily classify sentences as meaningful or nonsense.

Sentence Id	Sentence	Article
4	*"It is a county in the U.S. state of North Carolina"*	Anson County
5	*"the level experience is needed to level"*	Diablo II
6	*"Chess is a very complex game"*	Chess

Sentence Id	Times known	$P_\gamma(s)$	p-value	Degree assigned by scale	**Known**
4	9	0	0.0039	0.002	**No**
5	1	1	1	0.5	Unable
6	9	1	0.0039	0.998	**Yes**

Table 3.2: Examples of sentences played on the game. The sentence Id is used for reference in this chapter. Times known is the total number of times the sentence has been played (without counting nonsense votes). $P_\gamma(s)$ is the proportion of people that answered *true* or *false*. The p-value, $p_c(s)$, is the one obtained by the Binomial Hypothesis Test. The **degree assigned by our scale**, $\pi_c(s)$, represents the confidence we have regarding the decision to identify the sentence as known. The last column is the decision made regarding how known is the sentence with a significance of 0.1

Definition 4 Let $\pi_s(s)$ be the value that represents how much confidence we have on a sentence s being meaningful.

$$\pi_s(s) = \begin{cases} 1 - {}^{p_n(s)}/_2 & \text{if } P_\sigma(s) > 0.5 \\ {}^{p_n(s)}/_2 & \text{if } P_\sigma(s) \leq 0.5 \end{cases} \tag{3.7}$$

To classify the sentence as meaningful, we only need to define a threshold α against which we can compare $\pi_s(s)$. If $\pi_s(s) < \alpha$ we have a confidence of $1-\alpha$ that the sentence is nonsense, and if $\pi_s(s) > 1-\alpha$, we have a confidence of $1-\alpha$ that the sentence is meaningful. Otherwise, we can only conclude that we need more players to evaluate the sentence.

We perform a similar analysis to the one described previously to define a scale $\pi_c(s)$ that represents the fact that a given sentence s is commonly known. See Table 3.2.

In order to classify a sentence as commonsense we combine both $\pi_s(s)$ and $\pi_c(s)$.

Definition 5 Let $\pi(s)$ represent the confidence about a sentence being commonsense.

$$\pi(s) = \pi_s(s) \times \pi_c(s) \tag{3.8}$$

Table 3.3 shows the corresponding value of $\pi(s)$ of the sentences from Table 3.1. Notice that to classify a sentence as commonsense it requires both $\pi_s(s)$ and $\pi_c(s)$ to be high.

Sentence Id	Times Played	$\pi_s(s)$	$\pi_c(s)$	Commonsense Score	**Commonsense**
1	1	0.5	0.5	0.25	Unknown
2	6	0.9844	0.984	0.96899	**Yes**
3	6	0.016	0.5	0.00781	**No**
4	9	0.998	0.002	0.00195	**Domain-specific**
5	8	0.004	0.5	0.00195	**No**
6	9	0.998	0.998	0.9961	**Yes**

Table 3.3: The Sentence Id refers to the sentences in Tables 3.1 and 3.2. **Commonsense Score** is the score obtained by the sentence when evaluated for commonsense, $\pi(s)$. It represents the confidence that we have on identifying each sentence as commonsense. $\pi_s(s)$, and $\pi_c(s)$ are the scores obtained by the sentence when evaluated for *meaning* and *known*, respectively. Times Played is the total number of times the sentence has been played. The last column is the decision made regarding if the sentence is commonsense or not with a significance of 0.1

3.5.5 Results and Evaluation

The game was released on Facebook[2] and on a university website[3]. Within a period of five weeks, more than 150 people played the game and more than 3,000 sentences were evaluated.

We analyzed the coverage and reliability of the collected data, and identified the presence of knowledge that needs further classification. Our game offers an explicit way to detect knowledge that should be discarded due to errors or noise in the input of contributors. These features are achieved by the use of our scale $\pi(s)$. If a sentence is not nonsense, commonsense or domain-specific, then the game can be directed to present it to players more often until enough data has been collected to make a decision regarding such sentence.

Among all the similar systems reviewed, only LEARNER2 [Chklovski and Gil, 2005] (a system that collects commonsense knowledge in natural language from volunteer contributors) reports data about redundancy. Out of 6658 entries, only 2088 are different statements and 4416 entries yielded only 350 distinct statements. This means that they collected 1.29 entries per statement. These few entries per statement produce unreliable data, which means that only 350 statements can actually be trusted. In contrast, our game collected 6763 entries and generated 3011 evaluated sentences, with an average of 3.46 entries per statement. Therefore, our data is more reliable than that of LEARNER2. Figure 3.4 shows the comparison of coverage and reliability between LEARNER2 and our game.

[2]http://apps.facebook.com/turingrpg

[3]http://commonsense.cs.illinois.edu/turing_game/index.php

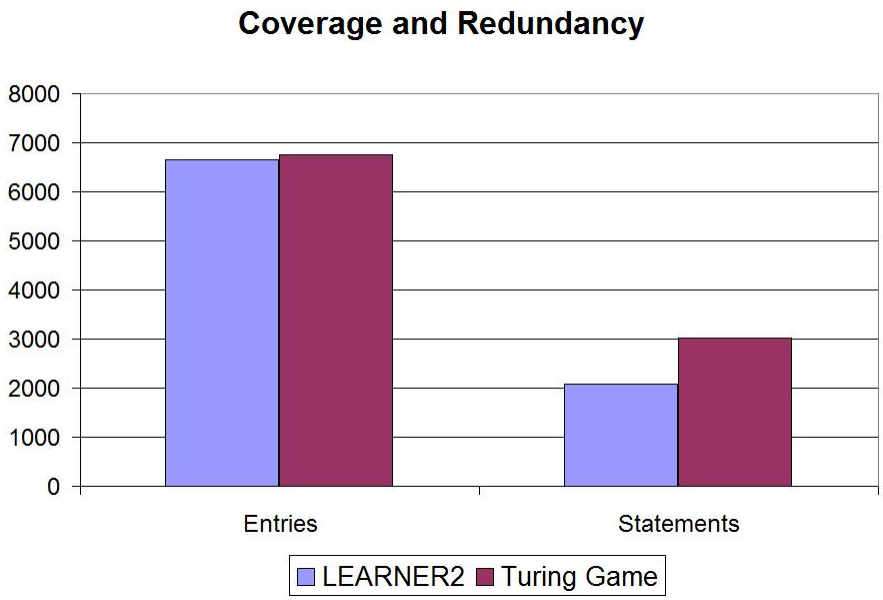


Figure 3.4: Comparison between LEARNER2 and the Turing Game in terms of coverage and reliability

For the evaluation, we asked 4 judges to classify a random sample of 50 sentences from our knowledge base. The judges evaluated the knowledge by classifying it in these categories: "Generally/Definitively True", "Sometimes/Probably True", "Unknown" and "Nonsense/Incomplete", which correspond to Commonsense, Domain-Specific, Unknown, and Nonsense, respectively. When comparing the answers of the judges to the ones from the game, the average agreement between players and judges was 94% ($\alpha = 0.1$).

In comparison to the other systems, Verbosity [von Ahn et al., 2006] asked the judges to rate each input as correct or incorrect; the judges reported 0.85 of the data to be correct, whereas LEARNER2 used a scale similar to ours and reported that 89.8% of the data that was entered by at least 2 people was correctly common knowledge.

3.5.6 Conclusions

In this section, we presented the design of a SSG that evaluates and classifies sentences extracted automatically from the Web. The main advantage of our design is that it classifies commonsense knowledge in a continuous scale, that measures how much the knowledge of a sentence is shared. which in turn, allows us to talk about how common a commonsense fact is. Our analysis of the results show that this SSG can be used to successfully

gather commonsense knowledge and allows for the empirical application of the definition of commonsense.

3.6 Application: Identification of Aggressive Individuals in Classrooms

The second application for SSG in this thesis relates to the problem of identifying aggressive individuals in classrooms (i.e., *bullies*). This problem is specially interesting to the framework of Social Sensing Games as it exploits all the areas of interest to SSGs: 1) designing an interface to collect relevant data, 2) infer offline attributes (roles) from online behavior, and 3) evaluate results using grounded theories.

3.6.1 Motivation

This application is inspired by the need of a new method for data collection in the social sciences that combines the power of lab-controlled experiments and the fine-grained observations of observational studies, and that enables direct testing of different social theories, such as RCT [Hawley et al., 2007].

We propose the design of a social sensing game that naturally enable examining social interactions under different social conditions and applies the SSG framework to study behaviors of adolescents. Our main objective is to show that the use of computer social games can increase the understanding of social relationships and interactions, in particular in the context of bullying. By using social computer games, scientists can take advantage of the increasing use of computers for managing social relationships while collecting large amounts of data that includes detailed information about social interactions and of the structure of the underlying social network of the participants.

The use of SSGs can also help to produce large amounts of data as the cost of running experiments and analyzing the data can be automated.

3.6.2 General Notions

The original goal for this game is to allow the answering of the following questions: 1) What are the friendship relationships amongst the participants, and what kind of interaction

do they have (cordial, aggressive, polite, etc.)? and 2) How are the loyalties and trust placed amongst the participants, and how do the rules of the game encourage leadership, competitiveness, etc.?

Therefore, the design of the game needs to be informed from social theories in order to guarantee that the data generated by the players is relevant to the problem (in this case, identification of aggressive individuals) and as such, the actions, tasks, and rewards need to be carefully selected.

After several sessions with experts in the field of Educational Psychology, it was decided that the game needs to emulate the circumstances of natural interactions amongst participants. Its features need to include the clear identity of the participants (in order to avoid anonymity and thus the online disinhibition effect), the presence of limited resources (i.e., points and coins), and the opportunity to collaborate and compete. The channel of communication also needs to be restricted to text messages, in order to have a non-intrusive way to monitor and analyze the interactions of participants with their peers (a method previously approved by an IRB). It was also decided that the game should have real world incentives besides in-game points in order to motivate engagement.

The main hypothesis is that there is going to be a subset of students (i.e., members of the classroom being studied) who are aggressive and that have a need for control and dominance, that will engage in coercive tactics directed toward non-friends, and that may solicit support for these tactics from friends within or outside their group.

Previous to the design of the game, ninety-six students from six different 5th grade classrooms in two Midwestern middle schools were administered three different surveys. These surveys were aimed at measuring aggression and delinquency[Espelage and Holt, 2001, Espelage et al., 2003]. The surveys included the following scales:

- The *Bully Scale*, which measures the frequency of teasing, name-calling, social exclusion, and rumor spreading.
- The *Fight Scale*, which measures the frequency of physical fighting.
- The *Victimization Scale*, which assesses verbal and physical peer victimization.
- The *Positive Attitude Toward Bullying and Willingness to Intervene*, which evaluates participants attitudes toward bullying, and the extent to which they are willing to assist a victim.

- The *Need for Control and Dominance*, which assesses self-perceptions of dominance and control within one's peer group.

Using the values of these scales, an expert labeled each participant as either a *bully* or a *non-bully* in order to have labels to evaluate the performance of the SSG.

3.6.3 Design as an SSG

Following the discussion above, the game was designed to be played in teams of 3 or 4 children belonging to the same classroom and that had been previously evaluated with the psychological surveys. These players were all between the ages of 10 and 12 years and had to answer a set of trivia questions. Also, participants had to nominate other participants with whom they would like to form a team, and the game then took them through a collaborative and a competitive stage. The output consisted of the team nominations and all the observed interactions between the players, including text messages and coins transactions.

With this in mind, we observe that the input X to the SSG has to be the social network of the classroom of the participants. Their attributes, relationships, and interactions are those that happen naturally in the classroom environment including demographics, friendships, rivalries, bully-victim relationship, talking, sharing, playing together, etc.

The game is to be able to capture some of the relationships between players by asking them about team preferences. The interactions are restricted to trading resources and communication through a chat interface. The actions available are then giving resources (in the form of coins), sending messages (with varied content) publicly or privately, nominating other players to be in the same team, and the task is to answer trivia questions while collaborating or competing. The reward is in the form points which are awarded if the instructions of each stage are followed. In order to make this points relevant, gift cards for the top scorer individual and the top scorer team were awarded at the end of the game. Formally,

Social Sensing Game for Identification of Aggressive Individuals in Classrooms

X =The classroom's social network

$o(X)$ =$A \times R \times I$, where

A ={`Demographics and social roles in the classrooms`}

R ={`friendships, rivalries, bully-victim relationships,...`}

I ={`Playing together`, `Talking`, `Fighting`, ...}

$Y = h(o(X))$ = Social roles of each of the participant (i.e., bully, non-bully)

$\mathcal{G} = \langle \mathcal{P}, \mathcal{A}, \mathcal{T}, \mathcal{R}_1, ..., \mathcal{R}_n \rangle$

$\mathcal{P}$ = Members of the classroom

$\mathcal{A}$ ={`Give coins`, `Send private message`, `Send public message`, `Nominate player positively`, `Nominate player negatively`}

$\mathcal{T}$ = Agree on trivia answer during collaborative stage and choose a losing player during competitive stage

$\mathcal{R}$ = Points are earned by following instructions during each stage and gift cards are awarded to the player and team who scores the most points.

Figure 3.5 shows screenshots of the nomination screen (top), a sample question of the collaborative stage (middle), and a sample question of the competitive stage (bottom). The game is played via a computer network and follows the steps described below:

1. Each user has the opportunity to nominate other participants whom they would or would not like to have on their team. In this stage of the game, we obtain information about task-directed peer nomination. Each team is currently created using a priori information gathered through surveys, ensuring that on each team there is at least one bully and one victim.

2. The second stage of the game consists on collaborating to answer a set of trivia questions, ensuring that all members of the team submit the same answer in order to obtain a reward (in this case, points in the form of coins).

3. The third stage is competitive or adversarial. During this task, each member of the team must provide a different answer to the question while one team member must choose a clearly wrong answer, effectively losing points while the rest of the team

wins points. In contrast to the collaborative task, in which the entire team must work together to maximize their individual reward, in the competitive task players are encouraged to directly oppose other members by convincing (or coercing) them to pick the wrong answer.

There are two winners in each game: a winning team (summing up all the individual rewards) and a winning player (the one with the largest amount of coins). During both tasks, participants are encouraged to use the chat system to coordinate their answers and to trade coins (points) amongst themselves.

During the collaborative stage of the game, team members work together to answer a set of 5 trivia questions (about topics such as history, geography, pop culture, etc.). Each question has four possible answers, and only one of them is correct. For each question, all the team members must agree on the right answer or otherwise no points are awarded to anyone in the team. These rules ensure that the players in the team must communicate and collaborate to agree on an answer, or everybody loses.

The fact that the team shares the payoffs and outcomes guarantees that everyone shares the same interest. For each question, team members have the option to peek at the correct answer in exchange for points from one of the players. In order to maximize the utility of the game, players must balance the peek penalty by sharing points amongst themselves. It is in the best interest of each player to share their knowledge about the question and not to let other players peek at the answers (in order to retain points).

During the competitive stage, each team receives a set of 5 trivia questions with four possible answers. Three of the four answers are correct and one of them is incorrect. In this task, each player on a team must choose a different answer for each question, with the added constraint that at least one member of the team must pick the wrong answer. Only the players that choose a correct answer get points. The wrong answer is marked explicitly (written in bold letters) in order to make it obvious and to encourage players to discuss who will pick such answer. It is in the best interest of each player not to pick the wrong answer, but also to ensure that someone else in their team picks it. This can only be accomplished by negotiating (either aggressively or non-aggressively) through the chat channel.

The output of the game consists of:

- The players' team preferences: friends/rivals nominations and the order in which they are selected.

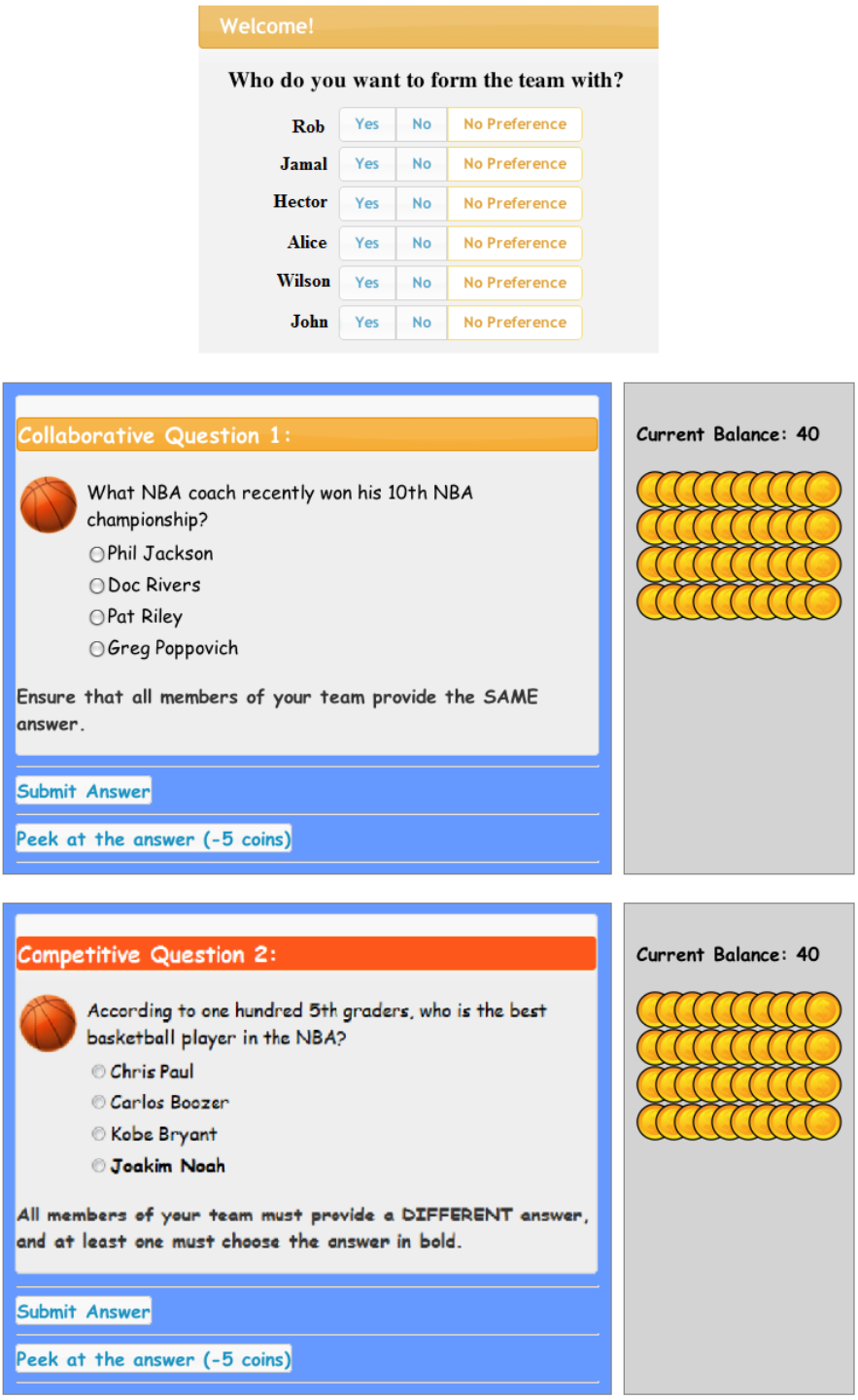


Figure 3.5: On the top: Nomination screen shot. Each player has the opportunity of stating with whom they want or don't want to play; or alternatively, of stating that they have no preference. On the middle: Screen shot of one question during the collaborative stage. On the bottom: Screen shot of one question during the competitive stage.

- Raw messages from users: all chat messages both in public and private channels along with the time at which the message was sent.
- Points transactions: transfer and forfeiture of points (e.g. in exchange for information).

3.6.4 Statistical Analysis of the Data

Using a 2(bully/non-bully) x 2(collaborative/competitive) ANOVA we studied the interactions and differences in the behaviors of Bullies and Non-Bullies (classified as such according to the data obtained through the surveys and analyzed by an expert) during both Competitive and Collaborative tasks. Results show that those two kinds of players behave differently during both tasks.

The features used for this analysis were (the abbreviated name of the features used in figures and tables is shown in parenthesis):

- The amount of private messages sent during the collaborative and the competitive stage (*prsent*).
- The number of private messages received (*prrec*).
- The number of public messages sent and received (*pusent* and *purec*).
- The number of times a player peeked at the answer (*peeked*).
- The number of points sent and received (*credsent* and *credrec*).
- Number of positive nominations sent and received, i.e., stating with how many people they want to play and how many want to play with them (*pnsent* and *pnrec*).
- Number of negative nominations sent and received, i.e., stating with how many people they don't want to play and how many do not want to play with them (*nnsent* and *nnrec*).
- Reciprocated nominations, i.e., number of people that nominated each other positively or negatively (*bpn* and *bnn*).
- Unreciprocated nominations, i.e., number of positive nominations to people that nominated the player negatively (*un*).

Table 3.4: Results of 2 x2 ANOVAs of Bully/Non-Bully and Collaborative/Competitive Stage. $^{+}$p<0.1, * p<0.05, **p<0.01

	prsent	*pusent*	*prrec*	*purec*	*credsent*	*credrec*	*peeked*
Bullies	23.12	20.69	19.79	50.86	2.57	3.5	1.55
Non-Bullies	15.77	17.93	16.71	50.82	3.39	3.25	1.06
p-value	0.037*	0.221	0.326	0.994	0.302	0.735	0.023*
	prsent	*pusent*	*prrec*	*purec*	*credsent*	*credrec*	*peeked*
Collaborative	20.18	22.35	20.22	61.62	3.98	4.01	2.01
Competitive	14.61	14.72	14.56	40.04	2.43	2.6	0.32
p-value	0.057^{+}	<0.001**	0.03*	<0.001**	0.019*	0.022*	<0.001**
	pnsent	*pnrec*	*nnsent*	*nnrec*	*bpn*	*bnn*	*un*
Bullies	6.19	6.05	5.05	3.71	3.05	1.33	1.33
Non-Bullies	6.02	5.97	3.31	3.79	3.26	1.08	1.18
p-value	0.791	0.925	0.09^{+}	0.914	0.671	0.566	0.685

Table 3.4 shows the average of the variables per type of player (i.e., Bully or Non-Bully), and the average per stage. There was a significant main effect of bully/non-bully on the amount of private messages sent (*prsent*), and the amounts of times peeked at the answer (*peeked*). Participants labeled as bullies sent more private messages and peeked at the answer a greater number of times than non-bullies. There was also a significant effect of bully/non-bully on the amount of negative nominations sent (*nnsent*); bullies sent more than non-bullies.

There was a significant main effect of collaborative/competitive on all variables. All players sent and received more messages (both public and private), sent and received more coins and peeked at the answer more during the collaborative stage. There were no significant interactions between bully/non-bully and collaborative/competitive. Taken together, the results suggest that bullies tend to send more private messages, to peek at answers more, and to send more negative nominations than non-bullies. In this analysis, the contents of the private messages was not analyzed but in the next chapter, the content is used to infer the player's off line role.

This analysis shows that the SSG is capturing quantitatively significant differences in the behavior of the players depending on whether they have the role of *bullies* or *non-bullies* in the classroom social network.

3.6.5 Qualitative Analysis of the Data

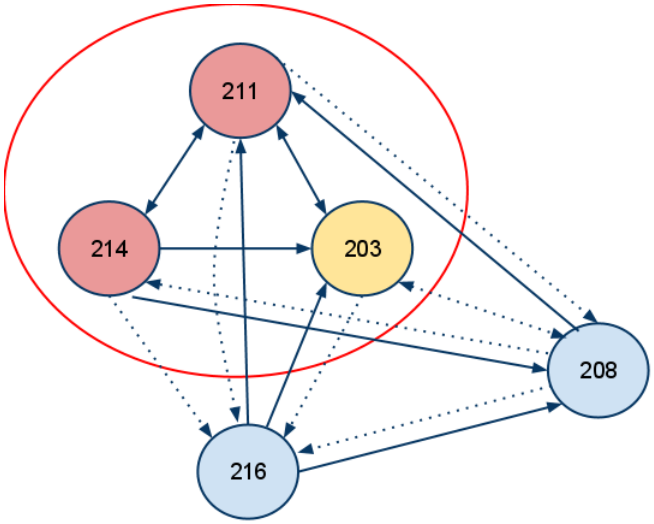
Further analysis also showed qualitative differences that are informative to social scientist that, to the best of our knowledge, were not available prior to the deployment of this SSG.

We present several examples which we consider specially interesting because they show behaviors relevant to bullying that cannot be detected using traditional research methods. In the following figures, solid arrows stand for positive nominations, while doted arrows for negative ones. Nodes colored red stand for participants labeled as *bullies* according to the survey, while blue nodes stand for *victims*. The red circle surrounding some of the nodes denote that they belong to the same team in the game.

The first example, shown in Figure 3.6, shows a subgraph of the nomination network of one classroom. According to the survey, the individuals labeled as *211*, *214*, and *203* are bullies, while *208* and *216* are victims. We can observe that participant *216* positively nominated *211*, *203* and *208*, but was negatively nominated by all of them. Notice that *211*, *203*, *208*, and *214* almost form a clique. By observing their chat messages, it is clear that *216* is experiencing some kind of victimization, but the survey cannot detect this type of pattern. It is also of special interest that *216* is being aggressive towards the others. Without going into the details of how the SSG generates labels yet, we can skip ahead and report that according to our inference algorithm all *211*, *214*, *203*, and *216* were labeled as *bullies* providing a discrepancy with previous measurements that points out possible shortcomings of the survey method.

Figures 3.7 and 3.8, show other examples of the output of the game. In Figure 3.7 we show the example of two participants, *104* and *108*, who are labeled as *bullies* according to their results on the surveys (but as *non-bullies* according to the game). Their interaction seems to be aggressive in nature but given their nominations we may be observing some type of friendly interaction between aggressive individuals as they did not seem aggressive towards others.

Figure 3.8 show another interesting example. Here we observe the interaction of two players in which one (*405*, a *bully* according to the survey) is trying to coax the another (*415*, a *victim*) into giving him his coins. *415* reacts by threatening *405* with soliciting assistance from two other players, *411* and *401*. We show the interaction of *405* with the other players which are, in turn, another *bully* (*411*, presumed friendly *415* but also friendly towards *405*) and *401* who has no label according to the surveys. *401* seems to be highly regarded by everyone and as we see, he is willing to pay for *415* in order to stop the harassment of *405*. Again, without going yet into the details of how these labels are generated, this participants

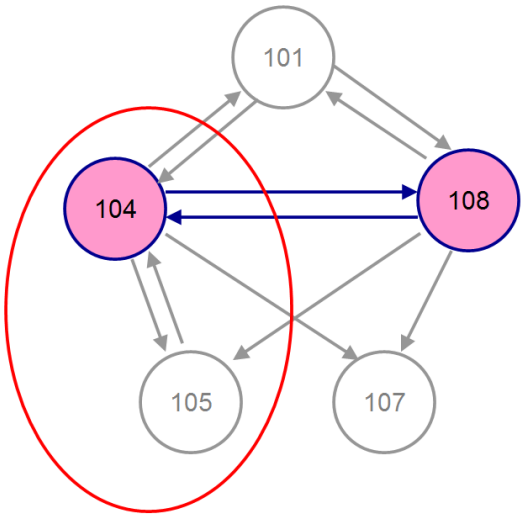


216 to 214: *send me some coins*
214 to 216: *no*
216 to 214: *dont talk to me then*

216 to 208: *send me some coins*
208 to 216: *no*
216 to 208: *why not*
208 to 216: *no*
216 to 208: *i said why*
208 to 216: *i hate you that why*
216 to 208: *f*** you*

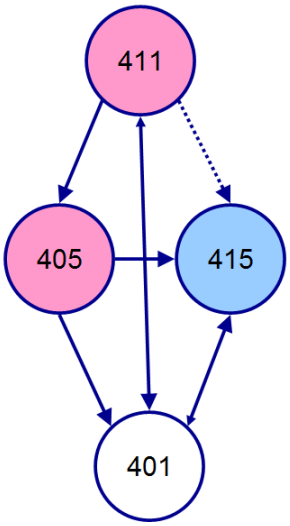
216 to 211: *hi*
211 to 216: *bi*
216 to 211: *shut up boy*
211 to 216: *you*

Figure 3.6: Example of victimization observed during game, but not captured by survey. Solid arrows show positive nominations, doted arrows show negative nominations. Chat messages show aggressiveness towards and from *216*.



104: WHAS UPPPPPPPP
108: lol
...
104: who on your team
108: shut up
...
104: GIVE ME COINS
108: no
104: SEND ME A COIN
104: FOR FAME MY WRATHHHHHH
108: nooooooooo
...
104: what ? u on
108: forget you
104: LOL
108: retard
104: IM NOT A RETARD
104: ALRUOGOGHG

Figure 3.7: Example of what may be consider friendly aggression in the game.



405 to 415: give me a coin
415 to 405: NO!!! LOL
405 to 415: why
415 to 405: do u want to get *411* on u again?
415 to 405: do you want me to get *401* on u?

401 to 405: stop bugging *415*
405 to 401: no, haa give me a coin and I well stop
401 to 405: no
405 to 401: yes
401 to 405: <sent coin>

411 to 405: u buging *415*
405 to 411: yes
411 to 405: haha she said u was u like her
405 to 411: no I need a coin
411 to 405: u want me to give u 1
405 to 411: yes
411 to 405: sike

Figure 3.8: Example of coercion and request for support inside the game.

were labeled by our inference algorithm as follows: *401* as a *bully* (this is probably an error from our algorithm), *405* correctly as a *bully*, *411* as a *non-bully* (this is a discrepancy from the surveys but consistent with the in-game observations), and finally *415* as a *bully*, this is also a discrepancy from the surveys but understandable as he coordinated what could be seen as retaliation inside the game.

3.7 Conclusions and Future Work

In this section, we have shown a Social Sensing Game that takes the social network of a classroom and generates, through measurements of gameplay, a heterogeneous social network (like the one shown in Figure 3.6) with data that is informative about the offline roles of the players.

The results (published in [Mancilla-Caceres et al., 2012b]) have shown that children labeled as bullies by educational psychologists play the game differently than non-bullies, i.e., they generate quantitatively different observations.

The ultimate goal of this research is to be able to create a tool that can be used broadly to help social scientists and educators better understand and prevent bullying. Currently, our system has shown that *bullies* and *non-bullies* (i.e., *bystanders* and *victims*) behave differently while playing the game, and therefore it can be used to identify bullies in classrooms that have not been surveyed yet. For this, a reliable model to identify individual bullies using only data from the game was developed and will be introduced in the next chapter.

Notice that by changing the design of the game, different information could be learned. For example, possible variations include: 1) changing the order of the tasks and determining how the change affects the way participants interact with one another; 2) having a finite number of points for an entire team and only changing the distribution of the points according to the performance in the game; and 3) forcing roles onto the members of the team to determine how quickly participants lead or follow others.

CHAPTER 4

INFERENCE FOR SOCIAL SENSING GAMES

In order to allow Social Sensing Games (SSG) to be useful for more than collecting data, it is necessary to be able to efficiently use the gathered information to learn or infer important information of the players of the SSG.

In this chapter we introduce an algorithm that takes as input the output of the SSG (i.e., a heterogeneous social network or HSN) and infers a label for each of the nodes in the network (e.g., assign labels to players of the game). One of the challenges in such social networks analysis is that the networks tend to be very large, and so, precise inference in them is intractable. Alternatively, traditional inference methods (such as ERG Models) assume the availability of large amounts of data that sometimes is not available. To deal with this challenges we need an algorithm that scales well in the presence of large amounts of data but that provides acceptable results when presented with limited information.

4.1 Motivation

Traditionally, social networks are represented as graphs, where either the nodes, the edges, or both are considered to be random variables. To analyze such networks, the most common approach is to use graph features such as degree of the nodes, number of dyads, number of cliques, etc., and to define a probability distribution over the network.

In other words, given a particular network, one is interested in obtaining the probability of observing that particular network. This probability can then be used to answer questions about the network such as predicting the evolution of the network, the characteristics of the nodes, and inferring the presence of an unobserved edge in the network.

The biggest challenge with this method is that in order to exactly compute the desired probability, it is required to consider all possible networks that can be observed in that particular scenario. This is usually a very large number that is in the order of $2^{O(n^2)}$ where n is the number of nodes in the graph. Furthermore, this number increases if we consider

networks that have more than one type of relations and nodes that have several types of attributes (i.e., *heterogeneous social networks*, the output of SSGs).

As an alternative to the method described above, we propose using strong assumptions about the types of edges in the network and perform inference in dyads of nodes. This method can reduce the complexity of the problem and can provide correct results as in some applications the complete structure of the network is only partially informative. For example, in the case of trying to identify aggressive individuals within classrooms, our results show that this assumption does not affect the results a lot [Mancilla-Caceres et al., 2012b].

This means that instead of analyzing the complete graph, it is possible to focus on pairs of nodes one at a time. In practice, this amounts to doing inference in many (less complex) models, which can be done more efficiently, and then merging all the predictions in order to obtain the desired results. If the assumptions described above hold, this algorithm would have a worst-case complexity $O(n^2)$ while still providing correct results.

This chapter will also provide some intuition about what kind of relationships can be analyzed with such algorithm and how can we characterize them. We do this by evaluating a heuristic that attempts to measure how transitive is a relation in a social network and use this measure to predict the performance of our algorithm.

4.2 General Notions

The main challenge that our proposed algorithm is trying to address is the fact that naïve analysis of social interactions might occlude all interactions (by assuming that all nodes in the network are independent of each other) or require large amounts of data by trying to include all dependencies.

For example, Consider an heterogeneous network of n nodes. Each node representing an individual that can interact with any other individual in m different ways. Also, assume that each individual i has a binary label y_i describing him.

Let $\overrightarrow{X}$ be the set of observations x_{ijk} representing whether interaction k was observed from individual i to individual j, and $\overrightarrow{Y}$ be the set of all labels y_i.

$$\begin{aligned} \overrightarrow{X} &= \{x_{111}, x_{112}, ..., x_{nnm}\} \\ \overrightarrow{Y} &= \{y_1, y_2, ..., y_n\} \end{aligned} \tag{4.1}$$

This means that there are up to 2^{n^2m} possible observations of the network and up to 2^n possible assignments to $\overrightarrow{Y}$.

Our goal is to estimate $\overrightarrow{Y}$ with a maximum-likelihood estimator (MLE) $\hat{Y}$ given the data $\overrightarrow{X}$. To do this, we will consider three possible estimators that make different assumptions.

The first one (named in this chapter *the classroom model* because of its usage in the SSG) makes no assumptions on the observed data and considers all possible observations to estimate all labels at once. This estimator is defined as:

$$\begin{aligned}\hat{Y} &= \arg\max_{\hat{Y}} P\left(\overrightarrow{Y} = \hat{Y} | \overrightarrow{X}\right) \\ &= \arg\max_{\hat{Y}} P\left(\overrightarrow{X} | \overrightarrow{Y} = \hat{Y}\right) P\left(\overrightarrow{Y} = \hat{Y}\right)\end{aligned} \tag{4.2}$$

Notice that because of the size of $\overrightarrow{X}$ and $\overrightarrow{Y}$, the likelihood $P\left(\overrightarrow{X} | \overrightarrow{Y} = \hat{Y}\right)$ requires 2^{n^2m+n} parameters to be learned from the data, where each observed network provides only 1 observation. Following our discussion at the beginning of this thesis that finding data relevant to certain social problems is hard (and therefore the need for SSGs), this is a big problem as, in practice, it is virtually impossible to learn this model appropriately.

The second alternative, named the *single-player model*, assumes that all individuals are independent of each other and aggregates all interactions across all players. Formally,

$$\begin{aligned}\hat{Y} &= \langle \hat{y}_1, \hat{y}_2, ..., \hat{y}_n \rangle, \text{ where} \\ \hat{y}_i &= \arg\max_{\hat{y}_i} P\left(y_i = \hat{y}_i | \mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{im}\right) \\ &= \arg\max_{\hat{y}_i} P\left(\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{im} | y_i = \hat{y}_i\right) P\left(y_i = \hat{y}_i\right), \text{ where} \\ \mathbf{x}_{ik} &= \sum_{j=1}^{n} x_{ijk}\end{aligned} \tag{4.3}$$

For this case, the likelihood requires us to learn only 2^m parameters from n different training points. If we assume $n \gg m$, this estimator is more feasible to learn in practice but it is likely very biased as the independence assumption between individuals is too strong and information about interactions is lost.

The third estimator (called the *pairwise model*) assumes independence only between each pair of interacting individuals, i.e., we assume that the way two people interact with one

Table 4.1: Comparison of three estimators to analyze the output of SSGs.

Model	Parameters to learn	Datapoints per network	Caveats
Classroom model	2^{n^2m+n}	1	It requires too much data.
Single-player model	2^m	n	Interactions are lost.
Pairwise model	2^m	n^2	Some context is preserved.

another is independent of how they interact with the rest. This is likely an assumption that usually does not hold but should produce an estimator less biased than the *single-player model* as some of the information about interactions is preserved.

Because the pairwise model analyzes each pairwise interaction, each player can generate up to n different labels that need to be aggregated according to some policy (e.g., by averaging). Formally,

$$\begin{aligned}
&\hat{Y} = \langle \hat{y}_1, \hat{y}_2, ..., \hat{y}_n \rangle \text{ , where} \\
&\hat{y}_i = \arg\max_{\hat{y}_i} \frac{1}{n} \sum_{j=1}^{n} P\left(y_i = \hat{y}_i | x_{ij1}, x_{ij2}, ..., x_{ijm}\right) \\
&\quad = \arg\max_{\hat{y}_i} P\left(x_{ij1}, x_{ij2}, ..., x_{ijm} | y_i = \hat{y}_i\right) P\left(y_i = \hat{y}_i\right)
\end{aligned} \tag{4.4}$$

Notice that this estimator requires 2^m parameters to be learned but, in this case, there are $O\left(n^2\right)$ interactions giving us more datapoints to learn a better estimator. Table 4.1 summarizes the three estimators and their properties.

In the rest of the chapter, we will present the *pairwise model* as an algorithm and show empirical results comparing the *single-player model* and the *pairwise model* for the case of the SSG for identifying aggressive individuals in the classroom.

4.3 Global Inference from Pairwise Interactions

From a formal point of view, the output of the SSG is an HSN Y that encodes the observed information from the real world. In most interesting cases, the size of the HSN and the amount of interactions is large enough that a naive treatment of the data can be prohibitive. We propose an algorithm that divides the HSN into dyads that can be analyzed efficiently using a variety of algorithms and that can be merged back to obtain a global answer for any

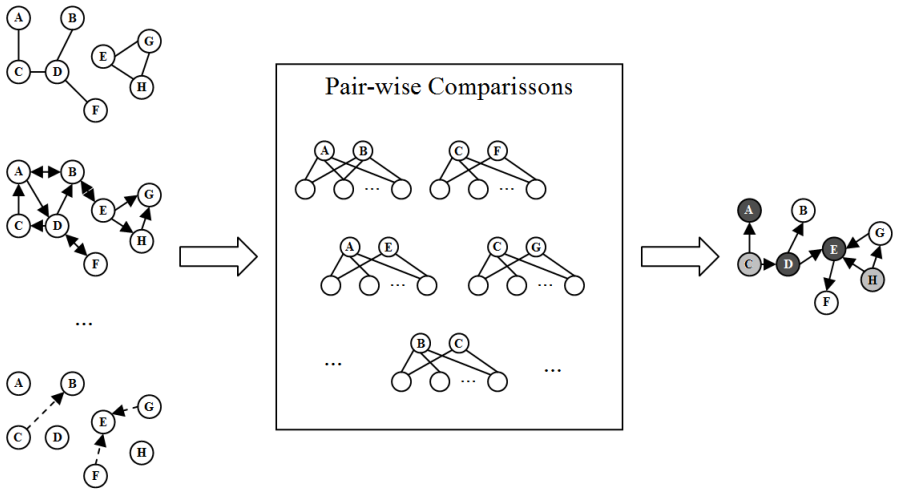


Figure 4.1: Inference algorithm. By dividing the network into dyads it is possible to infer labels for each of the nodes of the heterogeneous social network graph that serves as input.

question we want to answer. See Figure 4.1.

This algorithm is agnostic to the way in which we reason about each dyad and to how each of the results is merged to obtain a global answer. The strong assumption made by this algorithm is that each dyad is independent of one another, which in general is not true but in many cases can be safely assumed.

Let $Y = \{G_1, G_2, ..., G_m\}$ be an heterogeneous social network where $G_i = \langle V, E_i \rangle$ with V a set of nodes representing individuals, and E_i edges between nodes representing relationships (e.g., friendship) or interactions (e.g., phone calls) which may be weighted or not.

$$V = \{v_1, v_2, ..., v_n\} \tag{4.5}$$

where n is the number of individuals in the network.

$$E_i \subseteq \{\langle v_j, v_k \rangle\} \tag{4.6}$$

for all v_j and v_k that share a relation (or interaction) of type i. In order to handle weighted graphs, we can define a function w that takes a pair of nodes and a type of edge and assigns

a real value to the edge. Formally,

$$w : v_j, v_k, E_i \rightarrow \begin{cases} False & \text{if } E_i \text{ is unweighted and } \langle v_j, v_k \rangle \notin E_i \\ True & \text{if } E_i \text{ is unweighted and } \langle v_j, v_k \rangle \in E_i \\ r \in \mathbb{R} & \text{if } E_i \text{ is weighted} \end{cases} \tag{4.7}$$

Our procedure works by considering each pair of nodes that share at least one connection (i.e., share at least one edge) independently. For each pair, we can use the presence (or absence) of features to learn whether or not the target relation or attribute (i.e., the edge we want to predict) is present between the nodes of the pair. See Algorithm 1 for the pseudocode for this procedure.

Algorithm 1 Global inference from pairwise interactions.

Let $Y = \{G_1, G_2, ..., G_m\}$ be a heterogeneous social network where $G_i = \langle V, E_i \rangle$ with $V = \{v_1, v_2, ..., v_n\}$ and $E_i \subseteq \{\langle v_j, v_k \rangle\}$.

for v_j in V **do**
  for v_k in V **do**
    Create new pairwise model with v_j and v_k.
    for E_i in $\{E_1, E_2, ..., E_m\}$ **do**
      add feature with value $w(v_j, v_k, E_1)$ to pairwise model.
    end for
    Do inference with pairwise model (i.e., assign label to v_j and/or v_k).
  end for
  Merge pairwise prediction to generate global label for v_j.
end for
Return global labels.

This algorithm requires the specification of two important things at the moment of implementation. The first is the pairwise model used to characterize the interaction between pairs of nodes, and the second is the aggregation policy that is to be used to combine the predictions of the pairwise models.

Notice that this algorithm is in fact independent of the pairwise model and the aggregation policy and as such could be considered a family of algorithms. In the application shown later in this chapter, we will first argue why is important to use this algorithm of using pairwise interactions to do global inference and then we will specify and compare a particular pairwise model with other several alternatives.

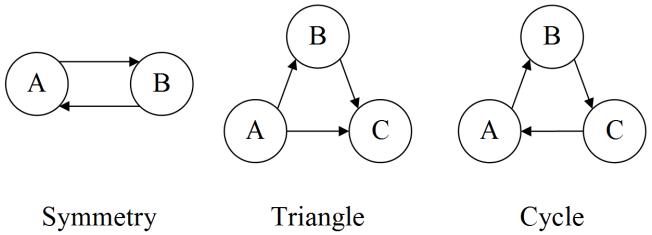


Figure 4.2: Examples of some of the dependencies that can be found in social networks. In this paper we propose that networks with low transitive dependencies (i.e., few triangles with respect to edges) can be analyzed with a pairwise model.

4.4 Limitations and Heuristics

The idea of dividing the network and analyzing it in smaller models requires the dependence of the edges of the network to be small or ideally, non-existent. Some types of networks may encode relationships that have this property whereas other do not. For example, the *friendship* relationship has been known to have the property of *homophily* [McPherson et al., 2001] (i.e., people who are similar like each other). What this implies is, for example, if Alice is a friend of Bob, and Bob is a friend of Claire, we expect Alice to be a friend of Claire. This is, of course, not always the case but it is a relatively common assumption to expect the presence of cliques in a network representing friendship. On the other hand, in a network composed of edges that signify *dating*, i.e., there is an edge between Alice and Bob if they have been in a date, we would expect to see very few cliques. For example, if we know that Alice has dated Bob, and that Bob has dated Claire, there is no reason to assume that Alice has dated Claire. These two examples refer only to one kind of dependence between edges, namely transitivity.

There are other types of dependencies that may arise in a network, e.g., symmetry (if Alice is connected to Bob, Bob is also connected to Alice), cycles (if Alice is connected to Bob, and Bob to Claire, Claire is also connected to Alice), stars, etc. (see Figure 4.2). The specific type of dependence that can (or is expected to) be observed in a network depends on the nature of the network. Traditionally, social network analysts use features that encode these dependencies. For example, a network might be represented by a vector of numbers representing the total number of edges, triangles, cycles, stars, etc. Then, a friendship network will probably have a large number of triangles or cycles, whereas a dating network will probably have a low number of these features.

Given the difficulty in performing exact (and sometimes even approximate) inference in

social networks represented with these features, we propose that by looking at the type of structures that appear in a particular network, we can estimate how strong these dependencies are. For example, if we observe that there are few triangles in comparison to the number of edges in the network, we can assume that most of the connections are not transitive, whereas if most of the edges are members of triangles, we can assume that transitivity is an important characteristic of the relation being represented.

Notice that any observed network is just a snapshot of the real-world social network. This means that at any particular moment, we are just observing a possible network, which also means that simply counting the total number of edges and the total number of cycles is not enough to guarantee that a relationship is transitive. What we propose (the proportion of cycles present in the network) is simply a heuristic that can be used to quantitatively determine whether or not it is reasonable to expect that the dependency between edges in a network is small. In particular, we will use such heuristic to account for the difference in the performance of our proposed algorithm when applied to a network with the purpose of predicting edges.

4.4.1 Heuristic for Estimating Transitivity in a Network

The heuristic that we propose, in order to evaluate the possibility that the edges are independent of one another, is the ratio of edges over triangles (i.e., undirected cycles and transitive relationships) present in the network, see Equation 4.8. In computing this heuristic, we make sure that the value of h is bounded between 0 and 1 by computing the ratio between the number of triangles present over the total number of possible triangles that can be encoded by a network with n nodes.

$$h = \frac{|triangles|}{(n-2) \times |edges|} \tag{4.8}$$

In the experiments below, we will show that in networks in which h is small, the accuracy of our algorithm is high in contrast to when h is large, regardless of the classifier used when predicting the labels.

4.4.2 Experiments

In order to test the hypothesis that the value of h is a good predictor of the accuracy of our algorithm, we ran it in the SocialEvolution dataset [Madan et al., 2012]. This dataset was collected in 2008 to track the everyday life of an undergraduate dormitory through mobile phones. It includes information about proximities, phone calls, and SMS messages as collected by the participants' mobile devices as well as survey results regarding health habits, music preferences, political beliefs, and relationships. In this study, we focused only on the information that relates to dyadic interactions in order to reduce the task to edge prediction, i.e., we focused on whether participants were physically close to each other (using the proximity data), whether they called or sent SMS messages to each other, whether they report each other as either friends, acquaintances or unknowns, and whether they know each other's music preferences.

Besides the fact that these relationships are dyadic, they were chosen among other possibilities in the dataset following the intuition that some of them will show high transitivity (e.g., friendship, proximity), whereas others will not (e.g., SMS or calling). For each of these relationships, we used the same algorithm to predict the edges and exactly computed the number of triangles and the number of edges for each case.

Notice that the learning task in this experiment is similar to the task of completing a dataset with missing information. In such cases, we have incomplete measurements about one particular feature and we can use the other available features to predict the missing one. For this particular dataset (and other similar ones) it would be difficult to complete the missing information using methods that make no assumptions on the dependencies of edges as we would need many several datasets of similar characteristics. Whereas if we assume independence between edges, we can use all the pairwise interactions independently and learn the most likely value for the missing edge.

In our case, we are interested in seeing whether or not the edges that we are trying to predict have a strong dependency among themselves. If they do, our algorithm is ignoring an important piece of information and will make many mistakes, whereas if the edges are mostly independent, our algorithm will have good performance.

4.4.3 Evaluation and Results

In order to simplify our experiments, we transformed the numerical values of the dataset (proximity, duration of calls, and total number of SMS messages) to binary values. Whenever

Table 4.2: Performance of algorithm with respect to classifier and target relation. Performance decreases as the heuristic h increases.

Target Relation	h	J48	Logistic	Naive Bayes
SMS	0.014	0.472	0.925	0.918
Phone Calls	0.019	0.549	0.763	0.763
Music Preference	0.036	0.497	0.761	0.76
Friendship	0.265	0.606	0.652	0.658
Proximity	0.289	0.597	0.607	0.604

a pair of participants was close to each other more times than the average pair, we assigned a value of 1 to the proximity feature and 0 otherwise. The same procedure was done with the total number of times participants called each other over the phone, and the total number of SMS messages sent and received. For the music preference feature, we compared the list of reported music genres liked by each participant and the list of music genres the others believed the first participant liked. If a participant mentioned at least one of the top three genres preferred by the other, we assigned a value of 1 and 0 otherwise. The relationship feature was originally nominal and we counted as friends those who mention each other as close friend or to socialize twice a week. In total we obtained 3732 pairs of interacting pairs.

For each of the 5 relationships, we ran our algorithm with 3 different classifiers while using the rest of variables as features. The three classifiers used were: J48 (decision tree), Naive Bayes, and Logistic Regression. For each experiment we evaluated the result using 10-fold cross validation and obtained the area under the ROC curve (AUC) as a measure of performance. We also computed the heuristic h for each of the features.

The results for each of the classifiers and each of the target relations are shown on Table 4.2. In two of the cases, with the exception J48 (see Figure 2), the performance behave as expected, i.e., at higher values of h the performance of the classifiers drops. For the case of J48, the classifier has a poor performance regardless of the value of the heuristic. When analyzing the contingency table for this case, we could see that the problem was that J48 was actually overfitting for most cases and was classifying everything with the label of the majority class. This is a problem of the algorithm itself and the fact that the classes were highly imbalanced. The other two algorithms did not show much difference in performance between themselves.

Figure 4.3 shows the results for each of the experiments with respect to the value of the heuristic. These results suggest that our intuition, that the value of h is a good predictor of accuracy for the performance of our algorithm, is correct. We can see that as h increases,

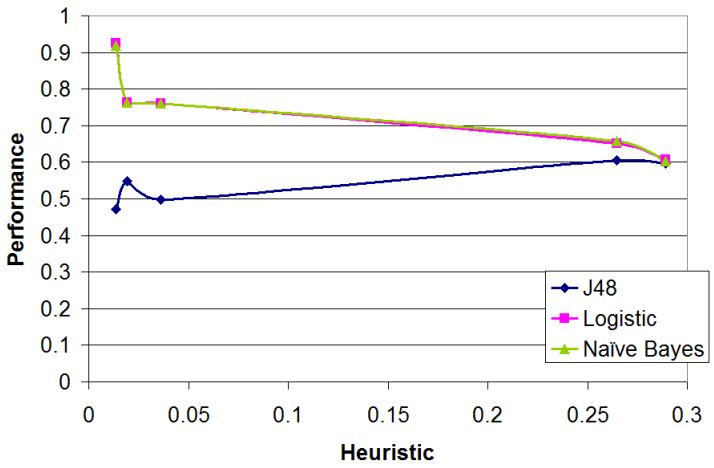


Figure 4.3: Performance vs. Heuristic (#cycles/(n-2)#edges). As the ratio of edges over cycles increases, the performance of the algorithm decreases regardless of which classifier is used.

the performance of our algorithm decreases with only one exception that has already been explained above.

In the figure, each line represents the performance of a particular classifier. From these results we can also hypothesize on the reason why the value of h changes for each target relation. As we can see, if we order the relations in increasing value of h we obtain: First, SMS messages with the lowest value of h, followed by phone calls, then knowledge of music preference, then friendship and finally, the relationship that measures physical proximity.

These results are very intuitive. As we discussed above, friendship is a type of relation in which we expect to see a lot of dependence between the edges. People tend to befriend people that are similar to themselves and therefore, it is very common to find cliques of friends which makes the value of h very high. For the case of proximity, it is also very reasonable that if one person is close to another which in turn is close to a third, the first one must also be close to the third. For the case of knowing the music preference of others, it is reasonable to expect that only people that know each other very well can guess the preferred music genre of each other which will be not necessarily transitive.

With respect to phone calls, even though we expect friends to call each other often, there are many other reasons for phone calls to occur. This makes it less likely that phone calls are a transitive relation. The same applies for the SMS messages, which tend to happen mostly in one direction (and possibly in the opposite direction) but not that much in cycles.

4.4.4 Conclusion

In this section we introduced a heuristic that predicts the performance of doing global inference with a pairwise model. This algorithm assumes that the target relation has no strong dependencies between edges and therefore it is possible to predict each edge by only considering features that involve the two nodes in question, avoiding the need to perform exact inference in the complete network (or having to approximate it). In cases where the target relation has low transitivity, the algorithm performed very well.

Our proposed heuristic counts the number of cycles over the total possible number of cycles in the network. This basically states that relationships that are not transitive should be easy to predict using our algorithm.

We performed an experiment involving real-world data and comparing the use of different classifiers within our algorithm and showed that in fact, whenever the target relation scores low in our heuristic, the performance of our algorithm is higher than when the relation scores high.

The value of this heuristic and algorithm is related to the problem of doing exact inference in social networks. If one can guarantee (or has strong reasons to believe) that a particular network has no strong dependencies between edges, our results show that it is not required to reason about the whole network together and it is not necessary to use other approximate inference methods, but simply to analyze each pair of nodes one at a time.

Because we are proposing a heuristic, the results shown in this section are by no means exhaustive. There are many different dependencies other than transitivity that may affect the performance of our algorithm. Nevertheless, the results show a clear correlation between the performance of the algorithm and the ratio of edges over triangles in the network.

There are other characteristics of the target relation that may affect the performance of our algorithm (e.g., sparsity, other types of dependencies such as symmetry, cycles, stars, etc.). It is not our intention to cover all possibilities, but to propose one particular heuristic that can be used to predict the performance of one particular algorithm.

In the future, we plan to further explore different heuristics that encode other dependencies and apply our framework to different datasets. Our goal is to better understand the type of networks than can be broken down into smaller models and simplify the work of social network analysts.

4.5 Application: Efficient Identification of Aggressive Individuals

The algorithm introduced in this section was used for the case of identification of aggressive individuals inside classrooms. Remember that, the HSN is formed by the players and by the in-game team nominations and interactions observed through the SSG, which include text messages and points transactions. By analyzing the interactions of each pair of players and averaging over all pairs, we are able to identify the role that each player has inside the classroom as either a bully or a non-bully.

The pairwise model used in this case was a two-layer Bayesian network (see Figure 4.4) that takes the interactions and nominations of every pair of players and generates a set of predictions for each one of them. Each prediction is the probability that a specific player behaves like a bully towards another player. The aggregation model for this application is the average of the results of the pairwise results. By looking at this average, we can generate a single label for each participant with relatively good accuracy and recall.

4.5.1 Pairwise Model

In order to properly create the pairwise model, we need to preprocess the output of the SSG. First, the chat messages in the game are unrestricted and written in natural language, which means that the messages must be classified into a small set in order to be able to create appropriate probability distributions.

We avoided the need of NLP techniques to classify the messages by having each message be labeled by 2 independent raters using 20 different categories describing the purpose of the message and its tone. The labels used by the raters are shown in Table 4.3 and example of the messages along with their labels are shown in Table 4.4.

We further classified them into 2 binary categories: Prosocial/Coercive messages, and Positive/Negative messages. This is because each player may interact through a private channel with any of the other players of the game and the messages may either constitute a prosocial message (e.g., helpful, agreeable, polite, etc.) or coercive (e.g., rude, aggressive, etc.). They may also express positive affect (e.g., happy, humorous, etc.) or negative affect (e.g., bored, sad, etc.).

As per our algorithm, we considered all pairwise interactions among players independently, i.e., for each pair of players that interacted during gameplay we gathered: the number of coins traded between players, the amount of messages sent by one of the players in the pair (the *sender*) to the other (the *recipient*), the nomination given by the *sender* to the

Table 4.3: Labels used by the raters to describe the text messages. Each message was assigned one type and up to two descriptors.

Types	Descriptors (Adjectives)	
1. Greeting	1. Friendly	15. Emphatic
2. Question	2. Polite	16. Casual
3. Request	3. Happy	17. Bored
4. Response	4. Helpful	18. Concerned
5. Offer	5. Playful	19. Unconfident
6. Command	6. Humorous	20. Aggressive
7. Accusation	7. Complimentary	21. Rude
8. Threat	8. Encouraging	22. Frustrated
9. Insult	9. Appreciative	23. Defensive
10. Statement	10. Agreeable	24. Disagreeable
11. Emoticon	11. Cooperative	25. Bargaining
12. Spam	12. Confirming	26. Argumentative
	13. Calm	27. Clarifying
	14. Confident	28. Random/Other

recipient (as either positive, negative, or don't care), and whether or not they belonged to the same team.

Figure 4.4 how an example of the pairwise model created for this application. It shows the case of two players (Adam and Bob) where Adam is the *sender* and Bob the *recipient*. It also shows the five binary variables observed during gameplay. These are *Wins*, *Coercive*, *Negative Affect*, *Nominated*, and *Same Team*.

The first one (*Wins*) takes the value of 1 if Adam got more coins from Bob taking into consideration all their interactions. The *coercive* and *negative affect* variables take the value

Table 4.4: Example of text message with labels assigned by raters.

Original Message	Rater #1	Rater #2
"be quite and get of my SITE"	Rude and Defensive Command	Aggressive Threat
"PEEK AT THE ANSWER"	Aggressive and Assertive Command	Assertive Command
"NO NO NO NO NO"	Frustrated and Assertive Response	Disagreeable and Argumentative Response
"cool!!!!"	Happy Statement	Playful Statement
"shut up david u dodo bird"	Rude and Aggressive Insult	Aggressive Command
"wat is the anaesr"	Casual Question	Aggressive Question

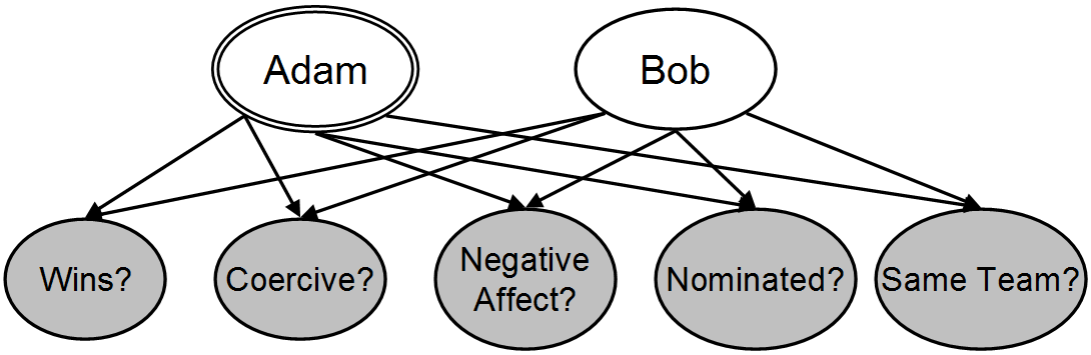


Figure 4.4: Example of two layer Bayesian network model for bully identification. By observing features readily available from the SSG we can infer a label for *Adam* given his interactions towards *Bob*.

of 1 if the majority of the text messages sent from Adam to Bob were labeled as coercive or expressing negative affect, respectively. The *nominated* variable has three possible values, 1 if Adam explicitly expressed that he wanted to play with Bob, -1 if Adam explicitly said that he did not want to play with Bob, and 0 otherwise. The *same team* variable takes the value of 1 if Adam and Bob belong to the same team and 0 otherwise.

This pairwise model is used to decide whether the *sender* (in the case of the example Adam) behaves in an aggressive fashion towards the *recipient* (in this case Bob). The basic assumption implied by our model is that the role of the two players, the *sender* and the *recipient* determine the kind of interaction they will have. For example, if the *sender* is bullying the *recipient*, it is to be expected that the amount of aggressive messages sent will be much larger than the ones received. In the same fashion, it is to be expected that the number of coins received and the nominations depend on the relationship of the players.

This model is created for each pair of players selecting each of them once as *sender* and as *recipient*. In our experiment, we were able to build 597 pairwise models and with those, we were able to compute the probability of any given *sender* to be a bully given the observation of the five features.

Handling Imbalance in the Data

Another of the challenges to address, in this particular dataset, is the intrinsic imbalance in the data; there tends to be a larger number of non-bullies than bullies in every classroom. In our dataset, the average number of bullies per class is 2, whereas the average size of a classroom is 15 students. That gives us a prior probability of 0.12 of being a bully. This

is a challenge because most off-the-shelf classifiers have an implicit decision threshold that assumes that both positive and negative classes are balanced, i.e., there is a 0.5 probability of a random example being a member of the positive class (in our case, of being a bully).

This implies that any observed interaction has a greater probability of coming from the interaction of two non-bullies (the majority class) than from any other possible combination (bully- non-bully, or bully-bully). This is exacerbated by the fact that interactions between bullies and non-bullies are not explicit, i.e., the way bullies interact with non-bullies is not necessarily salient. For example, finding a bully is not as easy as finding a player who says (or writes) more swear words, or who explicitly asks for money from another student. Therefore, interactions between the two most common roles (non-bullies) are more likely to be observed.

The data mining literature provides several ways to deal with this phenomenon (sometimes referred to as novelty detection). Among those suggestions are oversampling the minority class [Noto et al., 2008], changing the decision threshold whenever possible [Maloof, 2003] or redefining the negative and positive class to overcome the imbalance [Elkan and Noto, 2008]. In this work, we opted for changing the decision threshold due to the intuition that, in our case, it is appropriate to identify as *bully* those participants who have the highest probability of being a bully, regardless of the actual absolute value of the probability.

4.5.2 Training and Aggregation Policy

Even if a particular player was labeled as a bully using the survey, it does not imply that he will bully all the classmates with whom he/she interacts. For example, the fact that a bully may be coercive towards his victim, does not rule out the possibility of himself behaving nice and cooperative with the rest of people he interacts with (i.e., the player may be bi-strategic as predicted by RCT). Therefore, we are interested in calculating, for each training point (i.e., each pair of *sender-recipient*) the probability P_i^S of the *sender* S being a bully given his interaction with *recipient* i and the observed values of the rest of the variables.

$$P_i^S = P\left(S = \texttt{bully} | \overrightarrow{O}\right) = \frac{\sum_{r \in R} P\left(i = r, \overrightarrow{O} | S = \texttt{bully}\right) P\left(S = \texttt{bully}\right)}{P\left(\overrightarrow{O}\right)} \tag{4.9}$$

where $R = \{\texttt{bully}, \texttt{non-bully}\}$ and $\overrightarrow{O}$ is the observation vector corresponding to whether the *sender* won more coins against the *recipient*, sent more coercive messages, sent more negative messages, nominated the *recipient*, and whether or not they belonged to the same team.

This training procedure will generate n different probabilities for one specific *sender*, where n is the number of pairwise interactions of the *sender*. Because our survey data does not include information about pairwise interactions, we need to consolidate the n probabilities and obtain only one prediction per player. It is our intuition that non-bullies will have a low probability of being a bully across all the interactions they have, whereas a bully will have at least one interaction with high probability. Therefore, we estimate the probability of a *sender* S being a bully ($\bar{P}^S$) by computing the average over all the probabilities across all the *sender*'s interactions (See equation 4.10). It is expected that non-bullies will have a lower $\bar{P}^S$ than bullies.

$$\bar{P}^S = \frac{1}{n}\sum_i P_i^S \tag{4.10}$$

If we imagine the real-world scenario of using our game as a tool to identify bullies in a previously unseen classroom, we can see that the best method for evaluation is *leave-one-classroom-out*, i.e., train on five of the available classrooms and evaluate on the sixth one. This is done using as test one classroom at a time, in a cross-fold validation fashion.

After training the pairwise model and averaging all the probabilities for each *sender*, we look for the optimal decision threshold *thr* to label a particular player as a bully. In order to find this appropriate threshold, we searched for the threshold that maximized the F1 measure (the harmonic mean of accuracy *acc* and recall *rec*) during training and used that threshold to evaluate the results during testing (See Figure 4.5 for the ROC curve generated by our pairwise model and other alternative algorithms).

$$thr = \underset{thr\in(0,1)}{\operatorname{argmax}} \left(F1\left(thr\right)\right) \tag{4.11}$$

where $F1 = 2 \times acc\left(thr\right) \times rec\left(thr\right) / \left(acc\left(thr\right) + rec\left(thr\right)\right)$ and where the measures of performance are defined as follows:

$$acc(thr) = \frac{tp_{thr} + tn_{thr}}{tp_{thr} + fn_{thr} + tn_{thr} + fp_{thr}} \tag{4.12}$$

$$rec(thr) = \frac{tp_{thr}}{tp_{thr} + fn_{thr}} \tag{4.13}$$

where tp_{thr}, fn_{thr}, tn_{thr}, and fp_{thr} stand for true positive, false negatives, true negatives, and false positives (while using *thr* as threshold), respectively. To obtain these values we

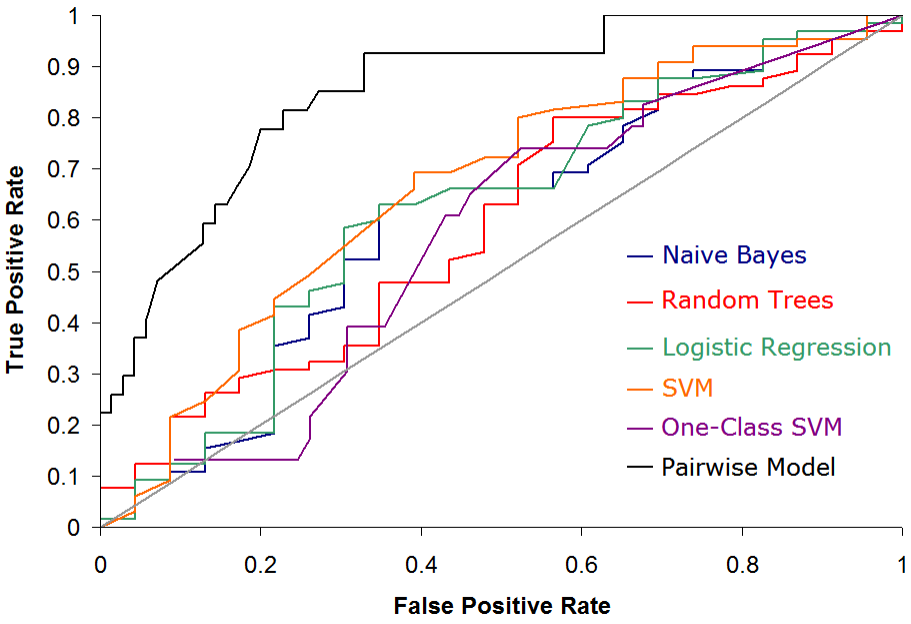


Figure 4.5: ROC Curve describing and comparing the performance of our pairwise model against other alternative models, when varying the decision threshold for labeling participants as *bullies* or *non-bullies*.

need to count how many *senders* have a probability $\bar{P}^S$ higher (or lower) than *thr* and were labeled using the surveys as either bullies or non-bullies.

That is, accuracy is the number of correct label assignments for both the positive and the negative class (*bully*, *non-bully*, respectively), and recall is the number of correctly identified bullies, from all the available bullies. Both measures of performance were used to keep in mind all possible scenarios in which our method might be used, and to overcome the shortcomings of each of these measures.

Accuracy, although probably the most common measure of performance, might be unreliable in cases where the data is imbalanced (as is ours) because a classifier might optimize this value by simply choosing to label all instances with the label of the majority, in our case, this would automatically obtain an accuracy of 0.88 which is the ratio of non-bullies in the dataset. Recall is also a very common measurement of performance, the problem in our case is that, because we are choosing the decision threshold during training, the optimization could always select the lowest possible threshold and label all players as bullies, which would ensure that all available bullies are correctly labeled, even though it is generating the maximum number of false positives. Because it is desirable that our system is capable of identifying all possible bullies in a classroom, and not only the ones who are *clearly* bullies,

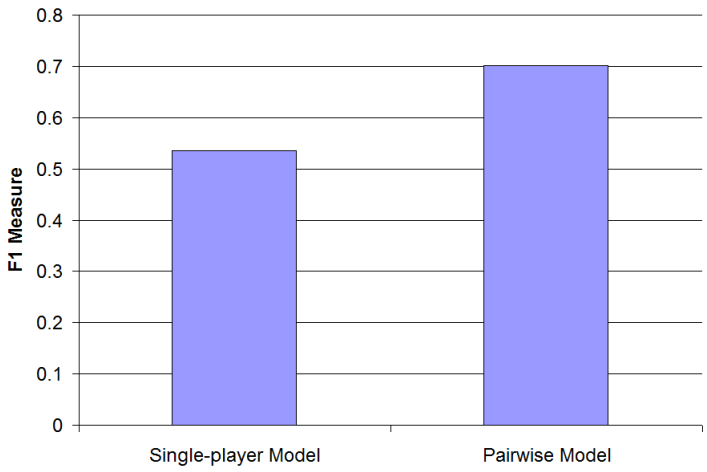


Figure 4.6: Comparison between performance of generating labels assuming all players are independent of each other and assuming pairwise independence.

the best decision threshold *thr*, is the one who maximizes both performance measures.

4.5.3 Evaluation

Once a threshold is found during training, the evaluation consists on obtaining the probability of a specific *sender* being a bully for each interacting pair in the test set, averaging the probabilities for each participant to obtain $\bar{P}^S$, and comparing this probability with the threshold. This allows us to obtain a prediction for each member of a classroom and compare it to the label generated by the surveys in order to obtain the values of accuracy and recall.

These *true* labels were generated using the score of the *bullyscale* of each participant and compared it to the average value of that scale (1.39 over 5) minus half the standard deviation (0.28). This means that the players that scored more that 1.11 on the bullyscale were considered *bullies* for the purposes of the evaluation.

For evaluation purposes, we first show the results of assuming independence of all the players, i.e., describing each player by aggregating the number of coins they received, the type of messages they sent and received, and the nominations they sent and received and trying to assign a label to each of them using off-the-shelf classifiers. The best results were given by using a Naïve Bayes classifier and are shown in Figure 4.6. When comparing this result with the best performance of the pairwise classifier (shown below), it can be seen that assuming pairwise interactions is better than assuming that all players are independent of

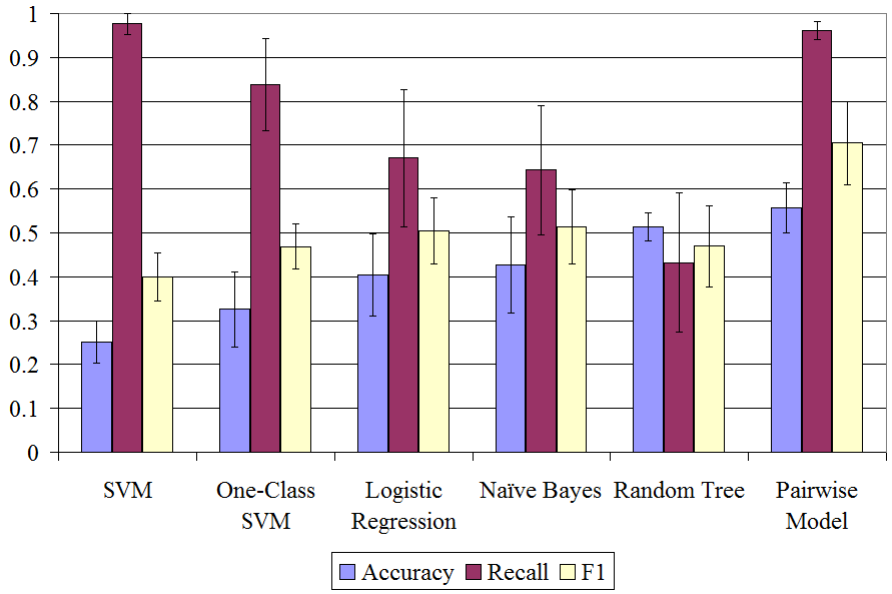


Figure 4.7: Comparison between off-the-shelf classifiers and our pairwise model, which accomplishes the best performance in terms of F1 measure.

each other.

After confirming that it is better to use a pairwise model, we compared the performance of our model with other five off-the-shelf classifiers that can serve as alternative pairwise models. These classifiers were ν-SVM [Schölkopf et al., 2000] (as implemented by LIBSVM [Chang and Lin, 2011] using a radial kernel), One-Class SVM [Schlkopf et al., 2001] (also, as implemented by LIBSVM), Logistic Regression, Random Trees and Naive Bayes Classifier (the last three as implemented on the WEKA Data Mining platform [Hall et al., 2009] with standard parameters). The results of these experiments are shown in Figure 4.7. The figure shows the average performance of each of the methods in terms of accuracy, recall, and F1 measure (along with the standard error as error bars) and shows how our pairwise model performs better than the other options for this particular application.

In order to compare the performance of each of the different classifiers on both accuracy and recall, a series of t-tests were performed to look for significant differences. The method with the highest recall is SVM (significantly better recall than Random Trees), and, as suspected, the cost of having an almost perfect recall is that of having a very low accuracy (it is actually the classifier with the smallest accuracy across all methods, i.e., SVM caused the threshold to be very low and therefore erroneously claiming that most, if not all, of

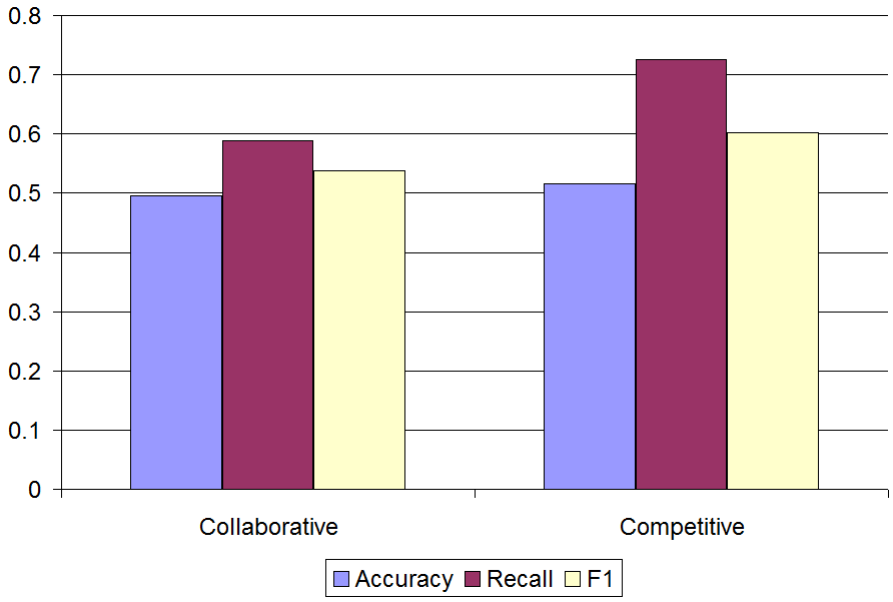


Figure 4.8: Comparison between our predictions using only the data from the collaborative and competitive stage.

the participants are bullies). In general, as methods improve on accuracy, the recall drops, except when using our pairwise model.

Statistically speaking, our pairwise model has the highest accuracy, which is significantly larger than the accuracy of all other methods, except that of Naive Bayes. In the case of recall, none of the methods (not even SVM) generates a statistically significantly larger recall than our model. This is due to the large variance of the recall on this method (actually, the only statistically significant difference in recall is given by SVM over Random Trees and One-Class SVM over Random Trees).

The main difference between all these pairwise models is the implicit assumption they have on the data. For example, SVM and One-Class SVM assume that the data is linearly separable (under a given kernel), the same goes for logistic regression. Random Forests impose no linearity assumption but may overfit the data for being to general. Naive Bayes assume independence between the features describing the pairwise interactions (whereas our model computes the exact probability given the features and assumes independence only on the interactions themselves and not on the features).

We also performed this analysis by dividing the data into the one produced during the Collaborative stage and the Competitive stage (under the intuition that the behavior on both

stages might be different and that the predictions may change accordingly). The results are shown in Figure 4.8. In terms of accuracy, there is no difference. In terms of recall, the performance of our analysis is slightly better during the competitive stage. Nevertheless, the results do not change by much than when the analysis uses the complete dataset.

4.5.4 Conclusions

In this section, we presented the analysis and results of applying an inference algorithm to the output of the SSG for identifying aggressive individuals in classrooms. The algorithm infers global labels from pairwise interactions by assuming that each interacting pairs of players is independent of the other. This assumption is very strong but the results seem to be reasonable given the available amount of data and the performance of other algorithms in the same task.

An important issue to address is the relatively low absolute values of both accuracy and recall, which in the case of our method are only slightly larger than 0.5. The most likely explanation for this is that either the model or the game is unable to effectively capture and identify the interaction patterns of all bullies in the game. Most likely, we are only capturing the patterns in the behavior of half the bullies. Although alarming at first sight, psychological research [Seigne et al., 2007, Monks et al., 2005, Naubuzoka, 2009] on bullying has found that there are two main types of bullies: those which have high executive functions (i.e., high capacity of negotiating such that they manage to get their way with both victims and teachers) and those with low executive functions (i.e., those with low capability of negotiating, get frustrated fast and disengage from socially accepted interactions). It is clear that this game would more easily detect those with high executive functions who manage to use the chat interface to manipulate their way into winning the game, whereas those with low executive functions will simply disengage from the game and stop contributing messages in the chat interface, making it extremely difficult for our game and model to identify them.

An alternative explanation is the data used to evaluate the algorithm, i.e., the "true" labels assigned by the experts. It turns out that most often, psychologists avoid the use of such strong labels and use more qualitative methods to understand the classroom's dynamics. In the next chapter, we address this by introducing two alternative analysis of the data: one through visualizations and the other through a purely statistical analysis of the game and survey data (without recurring to the labels of *bully* or *non-bully*) from a popularity and computer-mediated communication point of view.

CHAPTER 5

ALTERNATIVE ANALYSIS FOR SOCIAL SENSING GAMES

The application of our inference algorithm to the scenario of bullying identification (in the previous chapter) faces the difficulty that it requires labeled data. This is a problem as psychologists tend to avoid such strong labels that may create ethical issues and that, from the Computer Science perspective, requires extensive hand labeling.

For this reason, in this chapter, we explored the use of alternative methods of analysis, such as visualizations, as a way to better understand the relationship between pairwise interactions and global behavior and to highlight more qualitative patterns that may be hidden to our previous analysis.

Our intuition is that visualizations can increase the understanding of researchers about the social dynamics of the classroom while possibly affecting the interactions of the participants via self-introspection. To this end, we have developed a visualization tool comprising the available data, which generates no label for the participants of our SSG, but allows us to explore the interactions of each participant at the individual and classroom level.

Also, we explore an alternative quantitative analysis of the data generated by the SSG for identification of aggressive individuals by disregarding the labels of *bully* and *non-bully*, but instead focusing on the correlations between the data generated by the psychological surveys (i.e., the players' scores in the scales) and the data generated by the game (i.e., nominations, text messages, and trading of coins).

5.1 Visualization of Pairwise Interactions

In this section, we introduce *RelaVis*, a visualization tool that can be used to analyze and reflect around peer interactions to help researchers better understand the data from interfaces such as computer social games. This can validate the data collection tool itself, the previous survey results, and find new and interesting relationships between variables that may be missed or misunderstood when using less visual tools (e.g., simple linear correlation

analysis). For example, *RelaVis* has been presented to experts psychologists in the field of peer aggression and has highlighted possible pitfalls in previous analysis of data.

The objective of *RelaVis* is to allow researchers to explore social interaction and to highlight possible patterns. To this end, we created an interface with two main parts. The first (Figure 5.1 and 5.2) consists of a scatter plot where the researcher can choose any two variables recorded by the game (or entered by the researchers) and see their relationship. Users can also choose a third variable to see a ranking of the participants according to that variable.

By hovering over each dot in the scatter plot, or the name in the ranking, the researcher sees the name of the participants or their place in the scatter plot. The size of the marks in the plot represents the number of received nominations (a larger mark means more nominations received), and the shape of the marks represents gender (squares are used for males and circles for females).

Figure 5.1 shows two screenshots of the first view of *RelaVis*. The top image shows the relationship between the number of friendly messages received while playing and the score that each participant received in the survey measuring relational aggression. These two variables are significantly correlated with a coefficient of -0.26 (p-value = 0.0109). Nevertheless, the plot shows that the relationship between these two variables is not a simple linear one. Although it is clearly the case that people that receive more friendly messages score low on relational aggression (confirming the intuition that friendly behavior and aggression do not co occur), whenever a participant sends a small number of friendly messages (below 10 approximately) the distribution shows heterogeneity, i.e., there is possibly a moderator variable that can help us fully interpret data.

The bottom image shows the relationship between the number of friendly messages received and the score of the participant in the survey measuring caring behaviors. The correlation is now positive (0.232 with p-value = 0.02) and we see the heterogeneous nature of the data when participants send a small number of friendly messages, and a linear relationship when they send a large number (above 12) of messages.

These two figures highlight that participant behavior in the game supports the results obtained from self-report surveys when the behavior is clearly significant. For example, when players send many friendly messages they score high in prosocial behavior and score low in "antisocial" behavior. Also, we can see that in some cases there is a moderator variable that can help us explain the relationship between variables (such the number of friendly messages received and the scores in surveys) that requires further exploration.

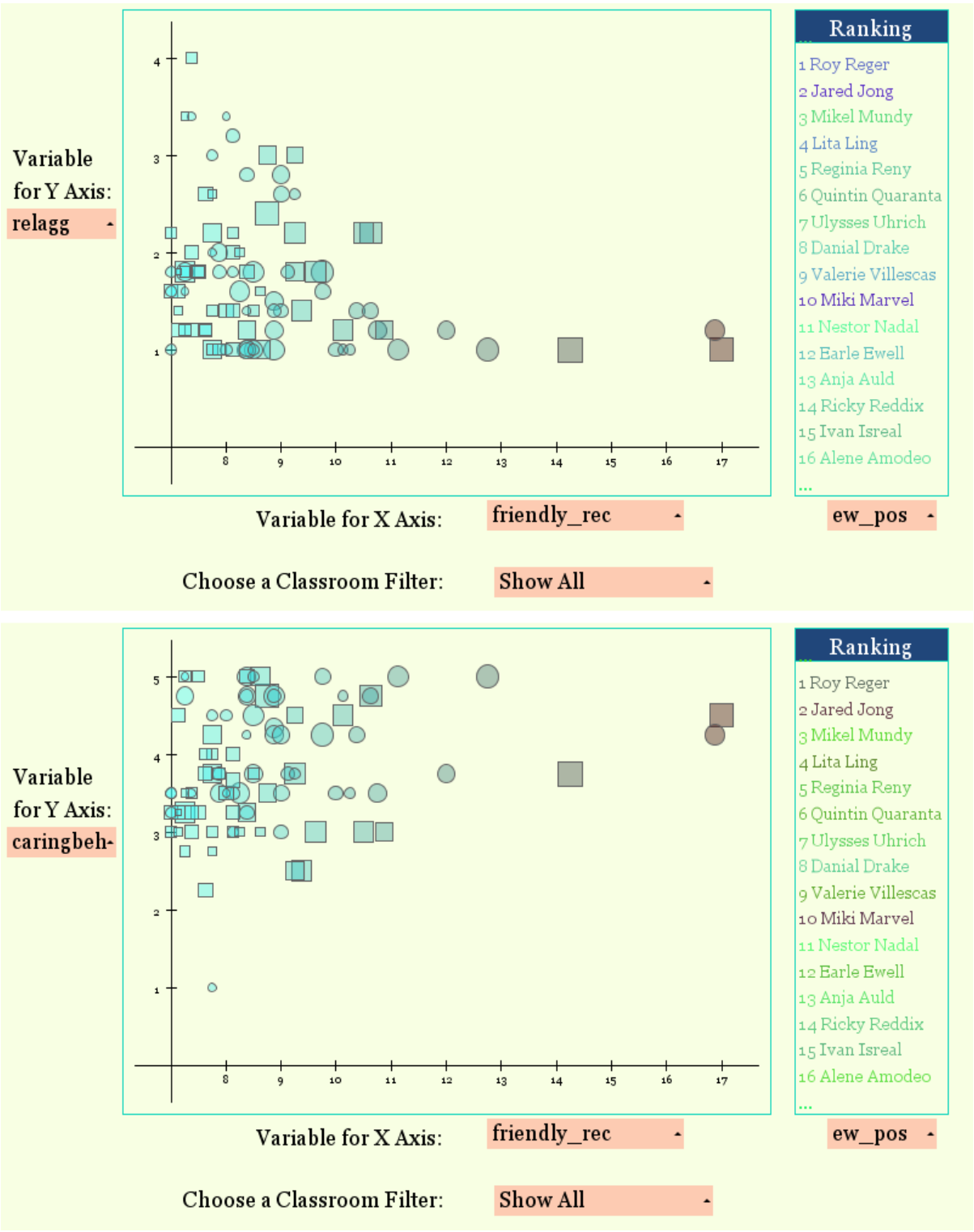


Figure 5.1: Screenshots of *RelaVis*. On the top, the relationship between relational aggression (*relagg*) and friendly messages received during gameplay (*friendly_rec*). On the bottom, the relationship between caring behaviors (*caringbeh*) and friendly messages received during gameplay.

Figure 5.2 shows examples of other relationships observed with the help of *RelaVis*. The top image shows the relationship between the number of rude messages sent by participants while playing the game and the score obtained in the scale measuring involvement in physical fights. Again, we see a heterogeneous distribution where most participants appear in the lower left corner of the plot. This is the desired behavior of participants as it implies low score on the fight scale and low number of rude messages sent, but it also shows that participants who send more than 10 rude messages score high in the fight scale.

The bottom image of Figure 5.2 shows a less clear relationship. It depicts the score of the victimization scale versus the final number of coins each participant had at the end of the game. The left part of the plot shows no relationship between victimization and the final number of coins, but the right side shows that participants with more than 10 coins have a decreasing probability of high victimization. While this is far from a linear relationship, it reveals that participants with the most points are not victimized, while losing the game does not allow us to infer anything about victimization.

Figure 5.3 shows the second *RelaVis* visualization. It allows researchers to graphically explore the pairwise interaction of participants. In this view, users of the tool can select any of the participants which then displays a list of the other participants with whom the selected participant interacted. It is then possible to observe a time line of interactions between the two participants (in the top) and the distribution of messages types sent. Hovering over the time line and the distribution it is possible to find more information about the particular message (its content) and the type. The time line provides a quick way to analyze the conversation between players while the distribution of messages allows for understanding the nature of the interactions. If the distribution is symmetric in the vertical direction, it implies both players sent and received similar message types. This helps to determine how aggressive was one participant towards the other. For example, in Figure 5.3 we see an interaction that differs both on the amount and type of messages exchanged by the players.

5.1.1 Conclusions

The use of a tool like *RelaVis* can help to explore the type of data gathered by the Social Sensing Games. It can help its users to gain new insights about the data collected and to possibly guide future research. It also allows the possibility to release or to show the data to interested parties in a way that certain points may be emphasized. This clearly opens the door to ethical issues that are addressed in the next chapter.

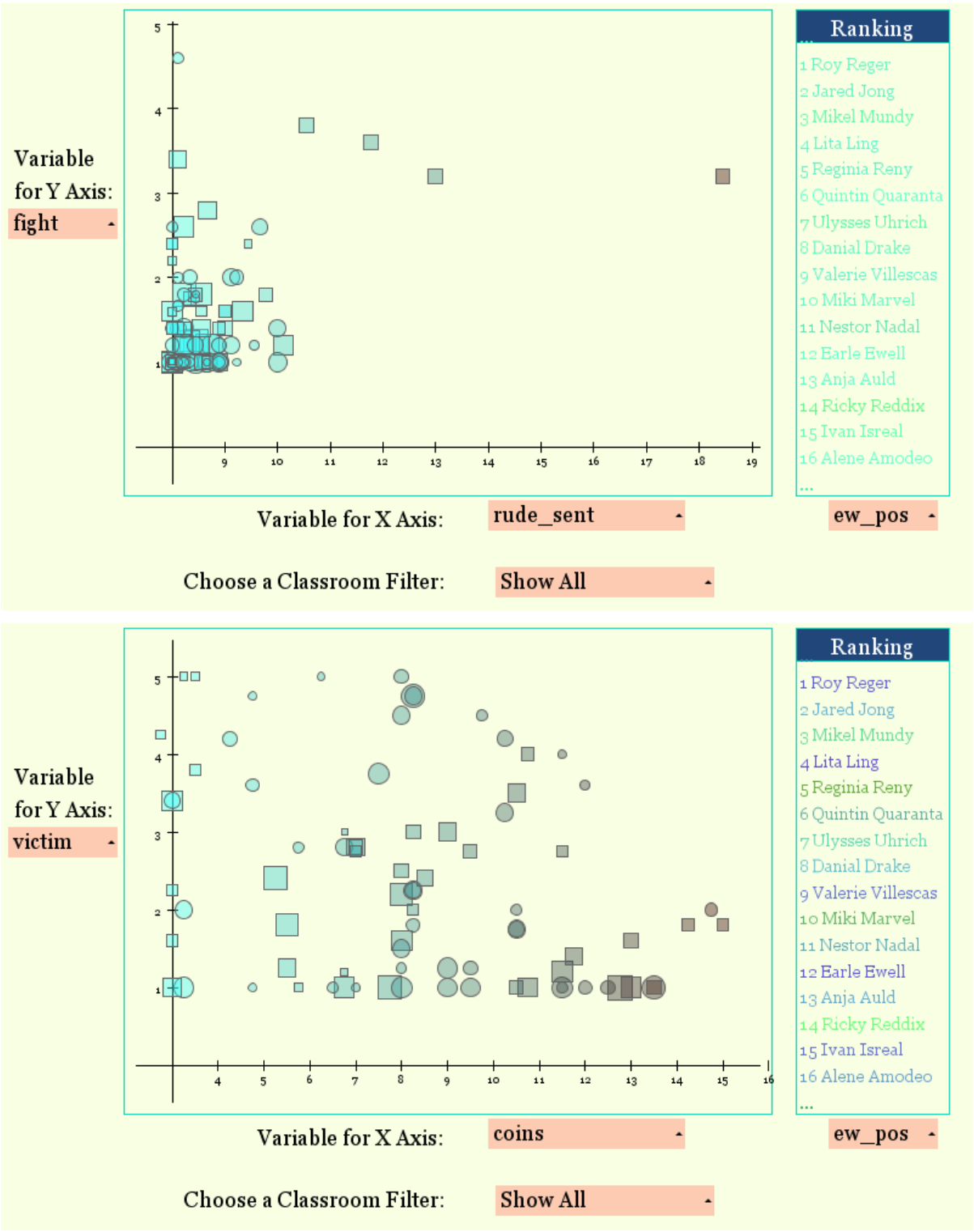


Figure 5.2: On the top, the relationship between the scores received in the *fight* scale and number of rude messages sent (*rude_sent*). On the bottom, relationship between score on the *victimization* scale and the final number of coins received during gameplay (*coins*).

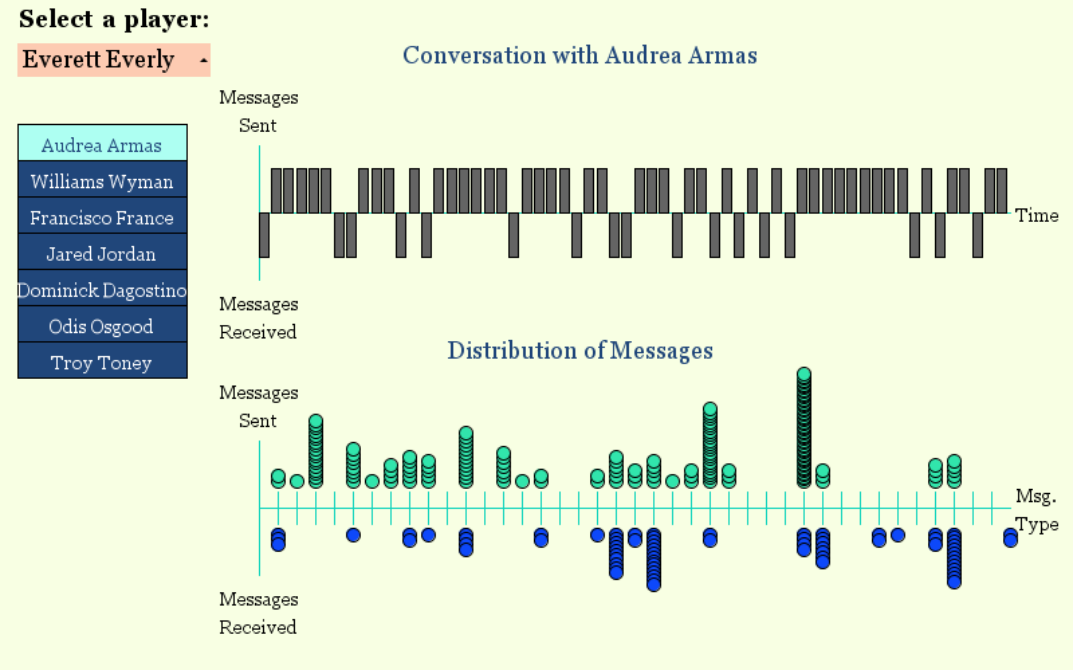


Figure 5.3: Screenshot of *RelaVis* showing the time line of the conversation between two players and the distribution over the type of messages.

5.2 Understanding Popularity and Computer-Mediated Communication

In this section, we present an alternative analysis of the SSG for identifying aggressive individuals. We analyze the survey data and the gameplay data without referencing the labels of *bully* and *non-bully*. We do this by exploring the way individuals are sought for forming teams and the impact of this selection on how the players interact among themselves. This can be considered a study on popularity and its implications in computer-mediated communication (CMC) [Mancilla-Caceres et al., 2013].

5.2.1 Motivation

Peer nomination is one of the main tools used by social scientists to study the structure of social networks. Traditionally, the nominations have been collected by asking participants to select a fixed number of peers, which in turn are all considered for the analysis with the same strength. In this section, we explore several different ways of measuring the popularity of

peers by taking into consideration not only the nominations themselves but their order and total quantity using the previously described SSG for identifying aggressive behavior. Notice that this analysis is possible thanks to the design of the SSG and its ability to collect fine-grained data about social interactions and nominations. It would probably be very difficult to do this kind of analysis with traditional survey methods.

5.2.2 Methodology

Using different metrics, we explore the relationship between the nominations and the players' interactions through text messages while playing the game. We propose five different metrics that can be used to find popular individuals among peers, which allow scientists to measure different characteristics of the individuals as shown by the correlations found between popularity scores and interaction variables.

The nominations and interactions between participants were collected through the SSG for identifying aggressive individuals [Mancilla-Caceres et al., 2012a]. The collected information forms a network of nominations that can be considered an indirect observation of the actual social network of the participants, while the interactions consisting of text messages provide information about the relationships between participants.

Our results support a new way of obtaining peer nominations by exploring their implications when paired with CMC, and show that the order and total amount of nominations can be used to have a more detailed analysis of peer interactions and CMC. Another contribution of our work is that the results show, contrary to previous results [Cairns et al., 1988], that aggressive individuals are not always rejected in peer nominations and that participants display both prosocial and aggressive behaviors, even toward peers that have been highly nominated. This suggests that there is value in interacting with aggressive individuals depending on the context (e.g., maybe they are stronger, better leaders, or more intelligent) and provides a justification for being aggressive while still being socially successful within the network. This can be explained by Resource Control Theory [Hawley, 2003] which suggests that aggressive individuals use both prosocial and coercive strategies in order to avoid the negative consequences of their aggressiveness while still being able to exploit some situations.

5.2.3 Popularity Metrics

We designed 5 different metrics of popularity based on the order and total amount of nominations, which provide different information and are defined as follows:

- *equal weights*: This simple method for handling nominations disregards the order and total number of nominations and assigns a weight of 1 to each nomination. It aims to represent how likable an individual is, because it indicates that the nominating player is OK to be in the same team as the nominee.

- *amount weights*: This metric normalizes the weight given to each nomination by the total number of nominations made by the nominating player, i.e., each nominated player receives a score of $1/n$ for each nomination, where n is the total number of nominations made by the nominating player. This is inspired by the intuition that if a player makes a lot of nominations, each nomination is less meaningful than if the player makes only one. This metric does not consider the order of nominations and we believe that it measures the strength of the tie or friendship.

- *order weights*: This metric takes into account the order in which the nominations are made. It assigns a score of $1/x$ where x is the position of the nomination, so the first nominated person gets a score of 1, the second a score of 1/2, the third 1/3, and so forth. This metric emphasizes the preferred nominations.

- *combined weight*: For this metric we combine the previous two and assign a weight of $\frac{1}{nx}$ where n is the total number of nominations made by the nominating player and x is the position of the nomination. A limitation of this metric is that a player that was nominated first in a long list of nominations receives a lower score than one who was nominated first in a short list, which may or may not be desirable.

- *inverse log weights*: This metric also combines both the total amount of nominations and the order in which the nominations are made. A player receives a score of $\frac{1}{1+n \log x}$ where n is the total number of nominations and x the position of the nomination. This function assigns a weight of 1 to the first nomination and a very low weight to later nominations in long lists. A player with a high score with this metric is one that is highly preferred to be in the same team by the nominating player.

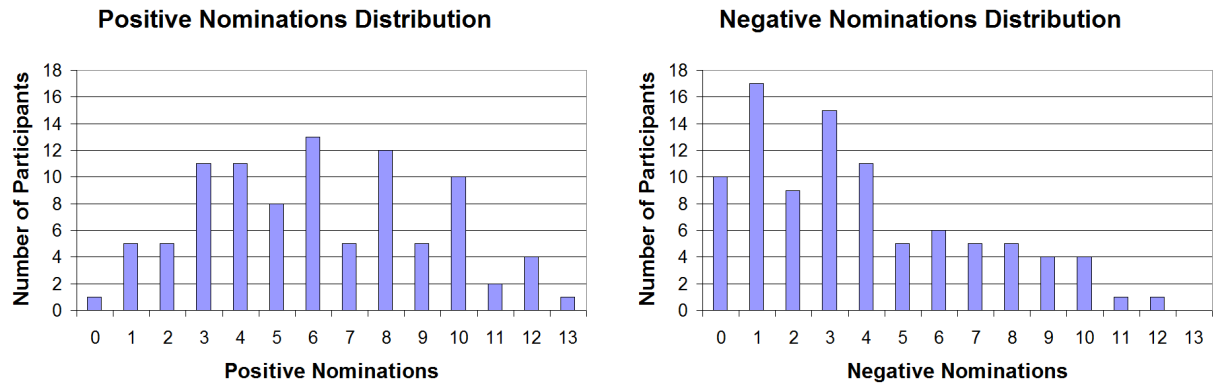


Figure 5.4: Distribution of positive (left) and negative (right) nominations. Positive nominations show a normal distribution, whereas negative nomitations show a skewed distribution.

5.2.4 Statistical Analysis

The popularity metrics were used in two different analyses. In the first one, a global positive and negative score was computed for each participant by summing all the weights generated by all the nominations received using all five metrics. Positive and negative nominations were computed independently because it is our intuition that these are orthogonal variables (i.e., a participant may be highly nominated positively and highly nominated negatively at the same time).

Figure 5.4 shows the distribution of positive and negative nominations. We observe a seemingly normal distribution of positive nominations, whereas the negative nominations show a skewed distribution, which indicates that most participants received at least few negative nominations and that few people received a large number of them. This supports our intuition that positive and negative nominations occur independently and therefore will be analyzed separately.

A correlational analysis was done with the popularity score, all the scores obtained through the surveys, and the variables collected from the interactions of the participants during gameplay. The second analysis focused on the pairwise interactions of each player. For every pair of players that shared at least one nomination we performed a correlational analysis of the respective score and all the variables collected from the interactions during gameplay between the two participants.

In the scenario where both players nominated each other, we analyzed both nominations independently, i.e., we explored what kind of interactions are seen when someone nominates another and when the other nominates the other one back.

Table 5.1: Most significant correlations between positive and negative nominations using 5 different metrics. The correlation coefficient is shown in parenthesis. The suffixes *.rec* and *.sent* stand for received and sent, respectively. * p-value < 0.05, ** $p - value < 0.01$.

	equal	amount	order
Positive nom.	private.rec (.49**)	private.sent (.39**)	private.rec (.23*)
	private.sent (.37**)	private.rec (.36**)	friendly.rec (.18**)
	caring behavior (.29**)	coins won (.29**)	happy.rec (.17**)
	friendly.rec (.25**)	caring behavior (.21*)	encouraging.rec (.13**)
	unsafe at school (-.25*)	unsafe at school (-.31**)	unsafe at school (-.27**)
	combined	log inverse	
	private.sent (.23*)	happy.rec (.14**)	
	argumentative.sent (.17**)	confident.rec (.12**)	
	happy.sent (.16**)	friendly.rec (.12**)	
	aggressive.sent (.15**)	encourage.rec (.12**)	
	unsafe at school (-.28**)	informal.rec (.10**)	
	equal	amount	order
Negative nom.	unsafe at school (.22*)	unsafe at school (.24*)	victimization (.23*)
	happy.rec (-.22**)	victimization (.21*)	happy.rec (-.11**)
	private.rec (-.25*)	coins won (-.22*)	confirm.rec (-.11**)
	coins won (-.26*)	private.rec (-.26*)	informal.rec (-.15**)
	informal.rec (-.28**)	informal.rec (-27**)	
	combined	log inverse	
	victimization (.23*)	victimization (.22*)	
	confirm.rec (-.10**)		
	happy.rec (-.11**)		
	informal.rec (-.14**)		

5.2.5 Global Analysis of Positive and Negative Nominations

The five strongest significant correlations found in the global analysis are shown in Table 5.1. In general, it was observed that positive nominations were highly correlated with prosocial attitudes (e.g., happy messages), particularly when using metrics that depend on the position of the nominations (*order*, *combined*, *log inverse*) and negatively correlated with unsafety at school. Negative nominations were correlated with feelings of unsafety at the school and with victimization, which suggests that participants who received a lot of positive nominations do not feel unsafe at school, whereas those with few positive nominations tend to feel unsafe.

For positive nominations measured through the *equal weight* metric we observe positive correlations with the amount of private messages sent and received. There was also a positive correlation with the score of the caring behavior survey suggesting that people that care about others tend to be nominated often (which supports our intuition that this metric mea-

sures likability of participants). This score was also positively correlated with receiving and sending messages with a prosocial tone, e.g., friendly messages. This metric was negatively correlated with the unsafety at school and relational aggression survey scores, showing again that *nice* people get a lot of nominations whereas aggressive people receive few.

When using the *amount weights* metric for positive nominations, we again observed a correlation with number of private messages sent and received. In contrast with the *equal weight* metric, we also observed a positive correlation with the number of coins obtained in the game, which suggests that when engaging in trade, participants with a high score in this metric tend to receive more coins than others. This supports our intuition that this metric measures how meaningful are the nominations, because people who were selected positively by themselves tend to have stronger ties with the players who nominated them and therefore receive more coins from them. As was the case with the *equal weight* metric, this popularity metric was also positively correlated with caring behaviors and negatively with unsafety at school.

With respect to the *order weights* metric, high positive scores are particularly positively correlated with prosocial messages sent and received. This again suggests closeness between the players and the people that they nominated first.

The positive nominations measured by the *combined weights* are notably positively correlated with argumentative and aggressive messages sent. This suggests that players who were only nominated first in short lists are either very close to the people who nominates them (and therefore can discuss and receive insults as jokes) or that such nominations are in reality not strong but very weak. This could happen in the scenario where the nominating player actually has no strong links in the classroom and simply selects a peer who "is not that bad".

When using the *inverse log weights* we only found a positive correlation with receiving prosocial messages. The fact that this metric penalizes all nominations strongly (except the first one and the second one in short lists) means that we are only finding the kinds of interactions that are common with the first nomination that is supposed to be the most meaningful. This suggests that this metric most likely characterizes the kind of communication between close friends, as there is also a correlation with the amount of informal messages sent.

Informal messages are those using some type of slang or informal construction of sentences, e.g., *"hows it goin"*, in contrast to a more polite construction such as *"mary, how are you doing?"*. This again suggests that people with a high score in the *log inverse weights* metric are more strongly connected to the people that nominated them and therefore are

Table 5.2: Example of significant correlations between sent and received nominations in pairs of players using 5 different metrics. At the right of each variable the correlation coefficient is shown in parenthesis. * p-value < 0.05, ** p-value< 0.01

equal	amount	order
question.send (.17**)	informal.send (.15**)	friendly.send (.19**)
informal.send (.17**)	informal.rec (.13**)	informal.rec (.15**)
friendly.send (.17**)	happy.send (.13**)	confident.send (.15**)
happy.send (.15**)	question.send (.12**)	helpful.send (.14**)
combined	**log inverse**	
appreciate.rec (.11*)	friendly.send (.17**)	
	confident.send (.16**)	
	humor.send (.14**)	
	helpful.send (.14**)	

closer and not required to be formal in their communication.

As mentioned above, the positive and negative nominations are independent of each other and tell us different things. In almost all metrics (with the exception of *equal weights*), having a high score in negative nominations is positively correlated with a high score in the victimization scale, which means that people that regard themselves as victims of the aggression of others do receive a high number of negative nominations. Also, negative nominations are negatively correlated with receiving prosocial messages (such as happy messages) and with receiving coins from others. This suggests that people with a lot of negative nominations are indeed targets of aggression or bullying victims and serve as validation of both the game and the surveys.

5.2.6 Pairwise Analysis of Metrics and Interactions

We also explored the correlations of interacting pairs of players and the respective nomination score given by each other. Table 5.2 shows the results.

Most scores are correlated with prosocial behavior. The main difference lies in the order in which the variables are correlated to high popularity scores. With the exception of the *combined weight* metric, all scores that depend on the order of the nominations (i.e., *order* and *log inverse*) emphasize friendliness, helpfulness and confidence; whereas the others (*equal* and *amount*) emphasize informality and questions sent. This suggests that, as expected, the order of the nominations are important when measuring friendliness, i.e., people who are selected first tend to receive more friendly messages and those who are selected late (regardless of

the fact the nomination might be positive) are not really "as positive" as those who received an early nomination.

Because different classrooms may have different dynamics, the interactions among friends and popular individuals might be different. Thus, we repeated the previous analyses for each classroom separately. That is, instead of using the aggregated data of all 96 participants, we divided them by classroom (six classrooms in total). Although in general, prosocial behavior was correlated with positive nominations, in some cases we observed interesting differences such as having both prosocial and coercive (aggressiveness and frustrated) messages positively correlated with high scores in nominations. This phenomenon could be explained using Resource Control Theory, which suggests that aggressive individuals (e.g., bullies) use both prosocial and coercive strategies. These observations could possibly be instances of bullies applying such strategies through the chat channels.

5.2.7 Conclusions

The network of nominations that is observed in the SSG does not encode friendship, understood as a mutual dyadic relationship, but popularity, i.e., the degree of acceptance by peers, which is related to willingness to play with the specific person. This popularity might be related to the expectation that the person will be a good addition to the team (i.e., the person is considered as someone with high problem-solving skills, but may or may not be a friend), to friendship (i.e., I don't care if the person is smart or not or if it will help me win the game, but I just want to play with him because I like him), to some feeling of aspiration (i.e., I will select the coolest guy in the classroom so I finally have a chance to interact with him and maybe earn his friendship) or simply to the fact that the person is the least bad of them all (i.e., I have no friends here but I have to choose someone so it might as well be him). This suggests that children might have a tendency to inappropriately include in their nominations others who have desirable characteristics and not only their friends. Our goal in this sense is to account for this phenomenon by comparing different metrics of popularity in our nomination network and relating it to the real interactions observed during gameplay. This, at the same time, allows us to understand the classroom dynamics, in particular in terms of aggression and victimization (which was the original goal of the SSG).

Our results show that high scores in nominations are usually correlated with prosocial interactions, in terms of the tone and type of messages sent among participants. The popularity score is also correlated with psychometric scores obtained through surveys that measure

Table 5.3: Number of instances of proactive aggressiveness and targets of proactive aggressiveness in the game. By proactive aggressive we refer to interactions of unprovoked aggression. We also include the results of the surveys for each of the players and the inferred label using the SSG described in Chapter 3.

Player ID	Proactive Instances	Target Instances	Bullyscale	Survey Label	SSG Label
501	10	1	1.25	*bully*	*bully*
504	7	3	NA	*non-bully*	*bully*
603	5	0	1.25	*bully*	*bully*
414	0	3	1	*non-bully*	*bully*
215	0	5	1.25	*bully*	*bully*
407	0	7	2	*bully*	*bully*

how caring participants are, how victimized they feel, and how unsafe they feel at school, among others.

As expected, participants who score high in popularity also score high in caring behaviors; whereas those who score low (either by having few positive nominations or a lot of negative ones) tend to feel victimized and unsafe.

The metrics proposed in this section are intended to measure different aspects of popularity, such as how likable is a participant, and how strong and meaningful are their relationships. Our results show different correlations of each metric with different types of interactions. Notably, all of them share positive correlations with prosocial messages.

The approach shown in this section is applicable in cases where someone has to choose teams or report on friendships for particular tasks, which makes our method useful for peer nominations research in the social sciences and in organizational settings where the network is expected to remain unchanged during the relevant period of time.

Future work could include the analysis the gender of the participants. Previous results [Bramoullé and Rogers, 2009] show that cross-gender interactions are related to popularity. Also, we plan to explore cultural differences in the way popularity and interactions occur in classrooms as suggested in [Keresteš and Milanović, 2006].

5.3 Analysis of Proactiveness in Aggression

In addition to the previous analysis. We also explored how proactive were the players in showing aggression in the game. Before discussing the results of the analysis, it is important to realize that neither the surveys (which are used for evaluation of the game) nor the game itself were designed to capture proactiveness and therefore the results shown here are not

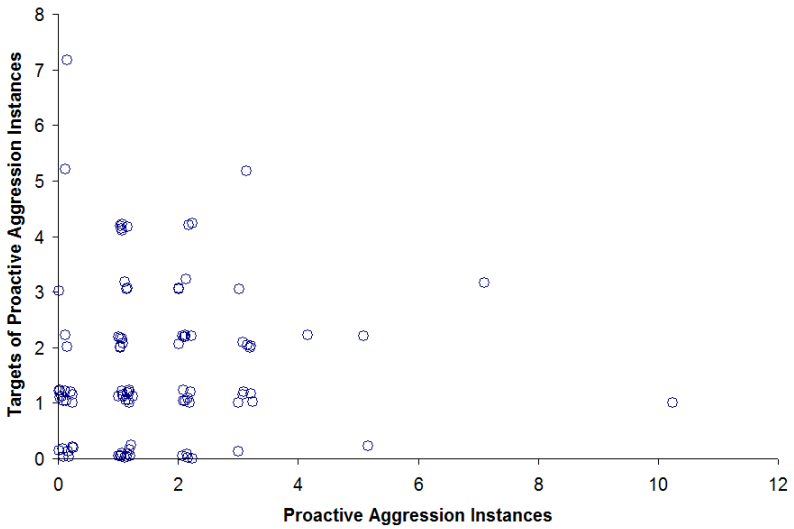


Figure 5.5: Distribution of proactive aggressiveness and targets of proactive aggressiveness.

conclusive and further studies would be needed to validate any conclusion.

In this analysis, we counted how many times a player started an unprovoked aggression with other players, i.e., for every other player with whom the first one interacted, we explored whether or not the first aggression was done by the first or the second player and labeled that interaction as *proactive* or *target* (to refer to the fact the the first player was a proactive aggressor or the target of a proactive aggressor, respectively).

Table 5.3 shows the results. We show only the top 3 proactive aggressors and the top 3 targets of proactive aggression along with their score on the bully scale of the surveys and the decision of the SSG. Notably, we see that the survey labels most of the players shown as *bullies* whereas the game labels them all as *bullies*. This highlights what we mention before (i.e., that neither the game nor the surveys account for proactiveness) but it also allows us to notice players that are clearly more proactive than others (or targets than others) that may need special attention by the researchers (again, without the need to use labels that fail to capture interesting informative behavior). As mentioned before, these results must be further supplemented with other studies in order to make better conclusions.

Figure 5.5 shows the distribution of proactive aggressors vs. targets. The purpose of this figure is to show how most players are close to the origin (i.e., are proactive aggressors or targets of proactive aggressors a small number of times). Although most players tend to be as much proactive aggressors as targets, there are some of them that are very skewed towards being only proactive or only targets which may be of special interest to researchers.

CHAPTER 6

PRIVACY AND ETHICAL CONCERNS OF SOCIAL SENSING GAMES

The emergence of new technologies and scientific methodologies help advance the current understanding of science and to create new research opportunities, but they also give rise to new ethical concerns [Bos et al., 2009]. In the case of Social Sensing Games (SSGs), they also introduce unforeseen ethical challenges related to privacy and anonymity, that need to be addressed.

One popular ethical position takes a consequentialist stance, which holds that what makes an action (or in our case, a design) right or wrong are its ultimate consequences [Light and McGrath, 2010]. From this perspective, the emphasis in technology design should be oriented towards achieving the desired goal (which should aim for the greatest good for the greatest number). In opposition to this stance is the user-oriented view that places technology in a morally neutral position and focuses the moral obligations and rights to the users of the technology. From this point of view, questions of morality do not arise as technology is regarded as neutral and the ethical decisions are delegated to the user. We subscribe to the first stance and analyze the risk and consequences of an inappropriate use of SSGs and what is the expected way of using them so as to minimize those risks.

With this in mind we wish to dedicate this chapter to the ethical implication of SSGs and their intended use. In particular, we will address three major question: What are the risks in terms of privacy and ethics of SSGs? What can be done to minimize these risks? and how are these risks related and handled by similar approaches to SSGs?

6.1 Ethical Risks and Privacy Concerns

Social Sensing Games collect information about the interaction of participants. These interactions include text messages which may include personal and private information. The gathered data may also include other type of information such as team preferences and results on tasks (that may have ill consequences for the participants). These are all sensitive

pieces of information that may violate ethical guidelines to the participants if released or analyzed in an inappropriate way (and if released publicly).

Besides the releasing of data, the way information is displayed and gathered during gameplay can be sensitive too (as well as any results obtained from it). The analysis done by the SSG may imply conclusions that can be harmful for the participants if not handled correctly or if the learned information is of a sensitive nature (e.g., labeling a kid a victim or a bully and releasing such information to teachers, parents, and the kids themselves).

As SSGs are somewhat similar to traditional observational studies (i.e., they facilitate the unobtrusive gathering of data), the potential harms associated with observational studies also apply to SSGs. Usually, these are generally less than with experimental studies, as no intrusive intervention takes place and participants are not subjects of experimentation (in other words the risk involved in taking part of our SSGs data collection is not larger than that of every day life). The most common risk in these cases are breaches of confidentiality, and as such, much of our security measures are oriented toward protecting the confidentiality of the participants.

There also exist risks that are proper to the application of SSGs (although not inherent to them). In the two applications explored in this thesis, different challenges and risk were observed. In the first application, the Turing Game (SSG for evaluating commonsense knowledge), the risks were minimum as no personal information was collected, but instead clicks were simply aggregated during gameplay. The only possible link that could be traced back to the participants identity was through a user id given by Facebook. The access to private information is then handled by Facebook policies regarding what Facebook Apps can and can't do. To further protect this information, in the final database this number was changed in a irreversible way. If the purpose of the Turing Game would have been another, personal information about the participants could have been accessed and different security measures might have been needed. A particular characteristic of this SSG that makes the risk minimum is that the only information collected was knowledge of particular random facts, which for completion we still protected but that provide no more risk than daily life activities.

The other application of SSGs shown in this thesis is the identification of aggressive individuals where the participants are children within the ages of 10 and 12 years old. This is a particularly sensitive population, that requires the consent of the parents or legal guardians and as such, data from such population requires special care. Therefore, the risks for this SSG were much larger and misuse of data is more dangerous. As much as the standard

techniques for protection were used, some of the literature proposes that these steps are not enough and greater care is necessary.

In our case, the SSG was used to study a population of children of particular middle-schools. Even though it is our opinion that it would be very difficult to identify the participants, we need to be aware that some apparently naive or harmless information can be used to breach the privacy of the participants (e.g., educated guesses of which middle-schools we might have worked with, references used in other papers from our collaborators that have strong relationships with certain schools, particular demographics, etc.). This again points out to the extreme security and consciousness with which the data that we collected should be treated.

The task of finding the identity of people through partial information available in datasets is usually called "social inference" [Motahari et al., 2009]. Social inference is possible due to the low entropy that exist in social situations or in some contexts. Given that people know some special people (from a particular demographic or location) it is relatively easy to narrow down the options to a small number of candidates.

For example, in the context of SSGs, if we were to show "anonymized" results to participants or teachers, it could be possible to narrow down and identify particular players due to the low entropy that may exist in a middle school classroom. This is particularly important where background knowledge exist, like in a classroom, e.g., it is very easy to know who just had a birthday, who is the only female Asian player, or who missed a couple of days of school.

Some of the harms that we can anticipate in case there is a breach of the data, misuse of the conclusions, social inference, or insensitive report of the analysis include: unfair labeling of the kids as bully/victims/bystanders, feelings of sadness/depression/anger by realizing that the self image of the personal role in the class is not as imagined (for example showing that no one wants to play with them, that everybody abuses them, etc), showing inadequate performance on tasks that may result in shame or public humiliation for the participant, generation of bias from other participants, teachers, or parents, among others. Some papers that also mention these risks on social inference and the importance of designing with ethics in mind are [Barth et al., 2006, Yakowitz, 2011]. All this means that the data must be securely protected and that all these possible harms must be considered before reporting or publishing the results.

Even though steps were taken to anonymize the data and to protect it from leakage. It is important to consider, and further study, better ways to protect the data from SSGs.

For example, recent analysis show the problems with simply anonymizing [Ohm, 2009, Zimmer, 2010a] as even some information (together with other freely available data or hints from the publications) can be used to re-identify people. Examples of this can be found in [L.Jedrzejczyk et al., 2009] and in [Zimmer, 2010b].

Part of the ethical risks of SSGs is that, contrary to survey research, we are collecting all type of information. This must be made clear both for the IRB and in the informed consent for the participants. Participants and researchers must realize that even though SSGs are similar to observational studies of public behavior, the expectancy of privacy made be compromised by the change of medium.

6.2 Security Measures

We will focus on three particular ways of minimizing the risks described above: protection of privacy (anonymization of data), informed consent, and restricted release of information.

As a first step, the risks described in the previous section imply that, at a minimum, the standard practices for protection of privacy should be followed. These include the anonynimization of the collected data (removal of names, restricted access to the data, etc.). Also, the data should be only accessed and shared (between participants or other audience) in an "as-needed" basis following the purposes of the study and the guidelines of the IRB.

With respect to the SSGs for identification of aggressive individuals, the kind of information collected and the procedure used to collect the data ensure that the participants are only partially identifiable. This is because participants were given a user id before accessing our SSG and the gathered data makes reference only to the user id (which is protected and accessible only to the researchers). Nevertheless, text messages may reveal information of the sender and recipients and could be used to identify and match the user id. To further protect the data, the user id (a number) and the name of the participant is kept in a different protected file and the names in the messages have been changed.

Once the data is gathered and anonymized, the kind of conclusions that can be obtained must be critically analyzed and studied as SSGs only capture a particular moment in the participants interactions, i.e., obtaining certain conclusions are risky and should be handled with care. For example, in the case for the SSG for identification of aggressive behavior, the conclusions and final labels assigned by the SSG should be verified with other type of information (such as standardized surveys) as it would be inappropriate to reach a definite conclusion given that the participants played the game only once (at least until this

application of SSGs has been more thoroughly analyzed).

Another important thing to consider is the fact that SSGs are being used to collect data from participants interactions in a non-obtrusive way. This means that participants act as naturally as possible (modulo that they are in a new situation and using a new tool). Nevertheless, SSGs pose similar intrusion as a video camera. Therefore, participants should be aware (as long as it does not compromises the study and the appropriate ethical guidelines are followed) of the function of the SSG and the fact that data has been collected. To address this, informed consent from the parents or guardians was obtained in the case of the SSG for bullying identification. This was not a concern for the SSG for evaluating commonsense knowledge and a general description of what data was collected was given to all participants.

Besides proper anonymization and care in avoiding risky results, special care to how and which information is shared with the participants is also necessary. SSGs have the possibility of being used as a way for self-reflection and as such, the kind of information that is provided (and how it is provided) may affect the participants in unexpected ways. As mentioned above, the realization of some participants about their social status within their peer group or the observation of how they victimize or are victimized by others may have unwanted effects in them. So far, we have addressed this risk by avoiding sharing the results with some of the interested parties until a better analysis is done and proper debriefings and procedures are in place.

As mentioned before, it is our stance that it is important to address the possible misuses of SSGs from the design. Depending on the application, the risks and potential harms that may come to participants of SSGs vary. Nevertheless, we subscribe to the consequentialist stance of ethics and therefore we believe that this dependence on the application does not exclude the ethical responsibility of SSGs for allowing the collection of such sensitive data. This concern must be addressed from the design of the SSG and its intended application so to ensure that the appropriate measures are taken for guaranteeing that the collected data will only be used as originally intended.

6.3 Security Measures in Similar Systems

As mentioned before, SSGs can serve as a data collection tool that has the power of obtaining data beyond what is necessary for a particular study, much like a powerful intrusive video camera. This means that many of the policies on privacy and ethics that apply to previous data collection methods (like observational studies) must be followed.

Also, due to the similarity to survey and experimental psychological studies and the current applications of SSGs, similar ethical principles and policies to such studies must be followed too. This includes, as previously mentioned, anonymization, and care with not inferring things about the data for which the experiment was not originally meant too, as many factors can affect the results and improper inferences may occur.

Among the common guidelines for ethical research are:

- Respect for people (and for their rights) which include autonomy. This is particularly important for people with impaired or diminished autonomy which fits for the case of the SSG for identifying aggressive individuals because our participants where children in a classroom activity.

- Justice, such that the burden of participation and the benefit is distributed equally. In the case of SSGs, this is specially important for the applications as the way that participants are selected (mainly through convenience) could affect the justice concern in the experiment.

- Beneficence and non-maleficence, which are of extreme importance as SSGs collect and use sensitive data. The risk of participating is low but privacy might be compromised and as such it is important to have extra care in the protection of data. We anonymized the data and restricted access to the data only by trained (and previously approved) research staff. Only aggregate results were published and conclusions of label generations are used only for statistics, not reported to any party outside the researchers. Some example of text messages were released in published papers but with no connection to the participants (only to a disconnected user id).

The literature gives us some examples of how to improve this kind of protections. The "Blackberry Project" (also known as the Friendship Project[1]) collects data that is similar to what an SSG would do, although SSGs are less intrusive, as data collection occurs only for a limited time and it stimulates only certain kind of communication. In a sense, SSGs are more similar to a lab experiment and observation study in the playground that the scenario that the Friendship project proposes.

For that project, the researches have taken common but thorough policies for protecting the data[2]. These include anonynimization of the data, extreme secure protection and

[1] http://www.utdallas.edu/ undrwd/index.htm

[2] see http://www.michaelzimmer.org/2012/04/25/research-ethics-and-the-blackberry-project/ for a review of them

restricted access to the data (by firms commonly protecting financial data), and recurring informed consents.

One particular potential risk of our SSG experiments, that is highlighted by the analysis of the security of the "Blackberry Project" is that parental respect for youth privacy and complete understanding of all the possible harmful implications of the results might not be sufficient and so the parental consent might not be appropriate. In order to deal with this situations we have refrained from releasing results of our analysis and publish only aggregated results. If we were to give results to teachers, parents might (in all their right) complain about the analysis of their children data and possibly breach of privacy. And even in the case that the results were not found to be a breach, the consequences that an automated process labels a kid could be harmful and questionable. Therefore, these results should be handled carefully.

6.4 Conclusions

The design and implementation of useful tools for research can follow ethical principles from the beginning of the design. From the point of view of technological design, researchers recognize the possibility of having an ethic-centric design [Friedman and Kahn, 2003] aimed to take into account human welfare, ownership, privacy, freedom of bias, universal usability, trust, autonomy, informed consent, accountability, identity, calmness and environmental sustainability. From the perspective of SSGs, and due to their similarity to classical psychological experiments and applications, in this section we focused mostly on their implications on privacy.

In conclusion, the design of SSGs need to be clear about the type of application that it is going to address and adhere to the strictest policies of ethical research and respect for privacy as their application requires (e.g., psychological studies IRBs, and other form of anonymization typical to online research).

CHAPTER 7

RELATED WORK

In this chapter we review some of the closest projects related to Social Sensing Games. First, let us restate that our work is inspired by the rise in popularity of computer-mediated communication and social interaction and by the increasing number of applications that are now open due to the reduced cost of collecting large amounts of relational data.

In this sense, SSGs are similar to other data collection strategies such as Human Computation (crowdsourcing, in particular) and to fields such as Gamification and Social Sensing as they use game concepts and are oriented towards the gathering of social information from the environment (see Chapter 2).

The main difference between Social Sensing Games and previous approaches that use games to solve difficult (computational) tasks is their goal and methodology. SSGs are not aimed at creating interfaces that encourage the participation of users to solve tasks, but to enable the observation of the ways in which participants interact with one another. In this sense, SSGs can be regarded as sensors that receive information from the world and create a simplified representation of the interactions and structure of the underlying social network.

In this chapter, we also include related work to visualization of relational information and to the areas of application of SSGs, i.e., collecting and evaluating commonsense knowledge and identification of aggressive individuals (or bullying).

7.1 Visualizations

There is a vast amount of research in the field of visualization, in particular that related to social visualization. Social visualization is a subfield of information visualization that attempts to make salient the connections and interactions of people for people [Karahalios and Viégas, 2006].

These types of visualization focus on varying social topics such as visualizing communities through conversations [Donath et al., 1999], visualizing individuals [Plaisant et al., 1996],

and studying information flow [Viégas et al., 2004], among others.

Closely related to the visualization tool employed to study SSGs, Leshed and colleagues [Leshed et al., 2009] explored visualizations of team conversations as a guidance for appropriate behavior. The main difference with our system is their audience and goal. Leshed's tool is aimed to teams of participants with the goal of improving their efficiency and participation, whereas our audience is the researchers studying peer interaction.

Also, Karahalios and Bergstrom [Bergstrom and Karahalios, 2009] introduce the concept of social mirrors as a type of social visualization aimed to users for perceiving themselves in real time and for exploration of group patterns and behaviors through visualizations. Although their objective is similar to ours, the main difference is their audience and methodology. They mostly focus on visualizing audio and conversations while we focus on text and other types of interactions such as team nominations.

Other studies use visualizations for different topics such as giving real-time visual feedback to groups [Tausczik and Pennebaker, 2013], triggering memories [Cosley et al., 2012], reflecting on memories [Isaacs et al., 2013], and exploring the implications and limitations of gathered data [Khovanskaya et al., 2013].

In [Isaacs et al., 2013], Isaacs and colleagues explored the use of technology for self-reflection about past feelings and actions to improve well being. *RelaVis* has not yet been evaluated in terms of self-reflection of the participants of our study, but the use of these visualizations by the scientist conducting the experiment can be considered a type of reflection on the experimental setup and can give unexpected insights about the social phenomenon at hand.

In more general terms, technology has been previously used to support health, in particular physical health (for example [Consolvo et al., 2009]) where the purpose is to encourage behavior change to improve a particular aspect of the individual health habits. Some work has also been done in the realm of psychological health where most of the work (see for example [Emmelkamp, 2005]) focuses on how technology can be part of a therapy program. This is not contrary to our final goal of using the visualization tool to help implement better interventions and to help participants reflect on their own behavior. The main difference is that currently, our work focuses on allowing better understanding of the participants social network mechanics as a way to move research forward rather than directly stimulating behavior change.

7.2 Work Related to Evaluating Commonsense Knowledge

Reasoning with commonsense knowledge has been one of the long standing goals of artificial intelligence since the beginnings of the field [McCarthy, 1968]. The SSG here presented focuses on the area of collecting and evaluating commonsense knowledge and as such, this literature review will focus on the best known attempts in this area of commonsense knowledge.

We start with the two best known (and probably the most successful) attempts to date: Cyc [Lenat et al., 1990] and OpenMind [Singh et al., 2002]. Cyc is a project that attempts to create a knowledge base containing all human commonsense knowledge. It uses a formal language developed specifically for Cyc, called CycL, based on predicate calculus and with a syntax similar to LISP (in order to represent knowledge in a formal way and to efficiently reason with it). Cyc employs knowledge engineers to manually code commonly known facts into different kind of contexts (e.g., *Bart Simpson* is a cartoon character in the real world, but in The Simpson's world he is a 9 year old kid) with the objective of overcome the brittleness observed in typical logic-based expert systems.

Some of the identified issues with this approach is that Cyc requires the user to be familiar with their language, and the fact that it requires people dedicated to input knowledge into the system in such a way that inconsistencies are not introduced (i.e., it shows scalability problems) makes it hard to use and scale. To address this issue, Cyc developed and maintained (for only a brief period of time) a game called FACTory used to input and evaluate contents of Cyc.

Another well known attempt to collect and evaluate commonsense knowledge was OpenMind. This project started on 1999 providing an online interface for allowing volunteers to enter unstructured knowledge into a knowledge base. Originally, it relied completely on the desired of participants to contribute (not unlike other well-known efforts such as Wikipedia) and as such it experienced a type of attrition. The project was continued through ConceptNet [Liu and Singh, 2004], a semantic graph containing all parsed knowledge of OpenMind and that continued to obtain knowledge through different crowdsourcing efforts and games with a purpose [von Ahn et al., 2006], as well as some automated methods [Scaiano, 2012].

Lastly, some other less-known efforts to collect common knowledge from contributors include LEARNER2 [Chklovski and Gil, 2005] which collected data about part-of relations through a kyosk from volunteers. Verbosity [von Ahn et al., 2006] was a GWAP that using the help of two players, filled in templates corresponding to commonsense knowledge (this data was later added to ConceptNet), and Common Consensus [Lieberman et al., 2007] that

asked players questions about how to achieve a given goal.

In summary, most previous approaches related to commonsense knowledge have focused on the collection of information (either from the web or from contributors). Very little effort exist on evaluating such knowledge (which was originally one of the contributions of the SSG *The Turing Game*) but, in particular, there has been no attempt to define commonsense knowledge beyond of *what everybody knows*. In contrast, our SSG departs from the assumption that commonsense is akin to a world model that is shared by people. In that sense, it is not enough to simply collect the data (and evaluate it) but to make it relevant to the people that generate it. Our SSG, by its design, names as commonsense those facts that are shared by the participants and keeps it relevant to those participants as well. This is something that can only be accomplished given the ability of SSGs to study the interactions and relationships between participants.

7.3 Work Related to Identifying Aggressive Individuals

Decades of sociological and psychological research have found that youth aggression and delinquency are strongly influenced by peer networks [Espelage et al., 2004]. Although psychometric tests and surveys have been rigorously designed by social scientists to collect and analyze information about social networks, such methods have the intrinsic limitation of being too costly to administer at a large scale and to relate the results to actual peer-to-peer interactions.

Even though the literature from psychology is vast, over the last few years there has been a large increase in awareness of the problem of relational aggression in youth (also referred to as bullying) [Lenhart et al., 2008, Microsoft, 2012, TheWhiteHouse, 2011], which in turn has inspired computer scientists to look for new ways to address such a problem [Dinakar et al., 2012, Xu et al., 2012].

In [Xu et al., 2012], the authors describe a system that analyzes twitter messages in order to identity the roles that occur in a bullying episode. These roles include that of a *bully*, *victim*, and *bystander*, as well as other less commonly identified such as *assistant*, *defender*, *reinforcer*, *reporter*, and *accuser*. The authors focus on the NLP algorithms required to find these traces of bullying in social media and on exploring the difficulty of determining the role given explicit messages. The task and the way the authors address this is related to sentiment analysis where they are interested not only in the polarity but also in the intention expressed in the message.

Another closely related study is presented in [Dinakar et al., 2012]. This work focuses on identifying implicit aggression in text messages on social networking sites, as such, it focuses on the detection and mitigation of cyberbullying, instead of face-to-face bullying. The main contribution of this paper relies on the use of commonsense knowledge to identify messages that traditional machine learning algorithms may miss. The authors also propose several intervention strategies that may be incorporated in their system including reflective user interfaces (e.g., notification or delaying of actions whenever aggression is detected), informing the user about hidden consequences (e.g., the approximate number of people that will see a particular message), and suggesting educational material. As in the case of Xu, et.al, our work differs from Dinakar's et. al. as ours focuses on learning about face-to-face bullying from online interactions, and not in the detection of implicit cyberbullying. Nevertheless, it is clear that our goals are similar and our SSG would benefit from the usage of such type of implicit detection and of their intervention strategies.

CHAPTER 8

CONCLUSIONS

8.1 Summary of Results

There are two major contributions brought forth by Social Sensing Games (SSGs) . First, we introduce a new method to collect relational data efficiently from people interactions and second, we present an algorithm to analyze the gathered data such that, when certain conditions are met, is efficient, correct, and scalable.

Through our case studies, we show that SSGs are aimed to applications that study social questions by showing promising results in the area of identifying aggressive individuals in classrooms (i.e., bullying) and in evaluating commonsense knowledge.

We also show the preliminary analysis of areas of expansion for SSGs that include the characterization of applications and networks that are suitable for SSGs as well as the usage of visualizations for better understanding and exploring the output of SSGs. We concluded with some exploration of some of the ethical risks that tools like SSGs introduce and some of the guidelines that should be followed.

One of our main goals at introducing Social Sensing Games is to allow social scientists to increase their understanding of social problems (such as *bullying* and *cyberbullying*) through the use of the data gathered by the game, the labels we infer, and the insights provided by the different analysis possible through visualizations and the correlational analysis.

This new tool is currently aimed to be used only by researchers in order to better understand the social problem at hand, i.e., the intended audience are scientists that uphold the ethical values described in this thesis and not the participants of the games or other interested parties that may use the results in unintended ways.

8.2 Conclusions and Contributions

Our results show that social games are a promising method to learn and infer social relationships between participants and that there are efficient algorithms to analyze such relationships, which in turn can help scientists learn information about the social network embedded in the game interactions.

In order for the social games to serve as sensors that provide meaningful data, it is necessary to design them in such a way that they capture the desired interactions. Taking into consideration that the design limits or defines what can be learned from the data and what data can be collected (through the definition of A,R,I,X,Y, and $\mathcal{G}$), this require a clear a priori specification of who the players will be, what actions can they take, what is the task to solve, and what kind of rewards are present in the game (either implicit or explicitly). Also, it is important to realize that the interface of the game, serving as a sensor, only maps part of the attributes, relationships, and interactions of the real world social network to the output of the SSG, expected to be a heterogeneous social network.

From the perspective of inference and learning. The fact that the output of the SSG is a heterogeneous social network, makes it important to have an algorithm that can handle large amounts of data (but at the same time can do well with few observations). Our results showed that by considering each pair of interacting agents (in our cases, players of the games) independently, we are able to obtain reasonable results without requiring a prohibitive amount of data collection while still providing scalability.

This method requires the assumption that interactions occur independently. This we acknowledge is a very strong assumption that only in some cases is reasonable to expect. We propose a simple heuristic that can be used to predict the performance of our algorithm which in turn can help minimize the cost of using this method in inappropriate cases. Our system does not restrict the use of better learning and inference algorithms and therefore, even though it would be ideal to develop better algorithms, the current stage of SSG does not include them but does not prevent them either.

8.3 Future Directions

This work gives rise to several possible expansions. Future directions for SSG may include a dynamic version of SSGs that allow for faster interventions as well as the exploration of alternative inference methods that generalize better (i.e., that put less restriction on the

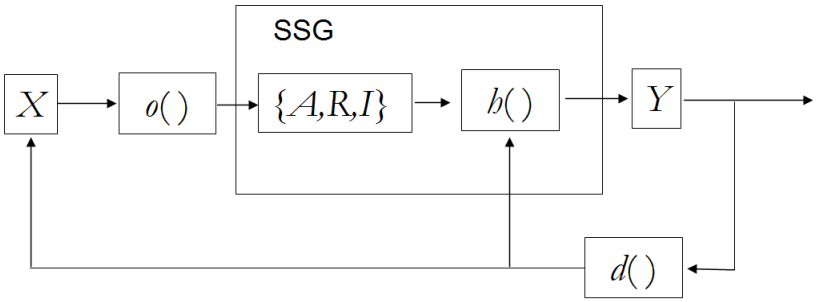


Figure 8.1: Adaptive Social Sensing Game. The proposed model suggests changing h as a function of Y (d in the figure).

characteristics of the social network being studied).

8.3.1 Dynamic Social Sensing Games

Currently, our games are only able to gather information from the participants and to generate labels for each one of them. In the case of the SSG for identification of aggressive individuals, we are interested in using such predictions in order to mitigate some of the factors that contribute to bullying. Psychological research [Garandeau et al., 2011] suggests that a common characteristic in classrooms afflicted by bullying is how hierarchical the social network is. Hierarchies can exist in terms of popularity, physical strength, or any other attribute.

Using the predictions from our learning algorithm, we should be able to infer the hierarchies within the classroom and adapt the rules and rewards of the game in order to alter its hierarchical structure. By allowing all participants to interact with one another in different contexts during gameplay, it is possible to promote the development of a better understanding and empathy towards one another through playing our game. Our intuition is that some relatively simple changes in the game structure can affect participants interactions. For example, changing the order of the tasks to solve in the game can affect whether participants cooperate or compete for the rest of the game. Also, having a finite number of resources (or points), or changing the teams during gameplay might help to promote more prosocial strategies. Other changes that we have in mind include: forcing players to have specific roles within the game (e.g., designate leaders), adding anonymity in some phases of the game, adding reflective interfaces that suggest different strategies for playing, and changing the reward structure for some players.

Figure 8.1 shows how adaptive games can be interpreted in the formal terms of SSGs. In the figure, we propose that by reading the current output of a SSG, it should be possible to obtain the pertinent change that needs to be done to h, i.e., how to change the set of actions, the task definition, or the rewards of the game, with the purpose of obtaining a different output Y. These changes would need to be informed from theory pertaining behavior change or learned from multiple simulations in order for the final change to occur in the real world X, as is desired.

REFERENCES

[Aboujaoude, 2012] Aboujaoude, E. (2012). *Virtually you: The dangerous powers of the e-personality*. WW Norton & Company.

[Aral and Walker, 2012] Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

[Bandura, 1986] Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ Prentice Hall.

[Banko and Etzioni, 2007] Banko, M. and Etzioni, O. (2007). Strategies for lifelong knowledge extraction from the web. In *K-CAP '07: Proceedings of the 4th international conference on Knowledge capture*, pages 95–102, New York, NY, USA. ACM.

[Barth et al., 2006] Barth, A., Datta, A., Mitchell, J. C., and Nissenbaum, H. (2006). Privacy and contextual integrity: Framework and applications. In *Proceedings of the 2006 IEEE Symposium on Security and Privacy*, SP '06, pages 184–198, Washington, DC, USA. IEEE Computer Society.

[Bergstrom and Karahalios, 2009] Bergstrom, T. and Karahalios, K. (2009). Social mirrors as social signals: transforming audio into graphics. *Computer Graphics and Applications, IEEE*, 29(5):22–32.

[Bernstein et al., 2009] Bernstein, M., Tan, D., Smith, G., Czerwinski, M., and Horvitz, E. (2009). Collabio: a game for annotating people within social networks. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, pages 97–100, New York, NY, USA. ACM.

[Bernstein et al., 2010] Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. (2010). Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 313–322. ACM.

[Blumenfeld, 2005] Blumenfeld, W. (2005). Cyberbullying: A new variation on an old theme. In *CHI 2005 Abuse Workshop, Portland, OR. Retrieved March*, volume 22, page 2007.

[Bos et al., 2009] Bos, N., Karahalios, K., Musgrove-Chvez, M., Poole, E. S., Thomas, J. C., and Yardi, S. (2009). Research ethics in the facebook era: privacy, anonymity, and oversight. In Jr., D. R. O., Arthur, R. B., Hinckley, K., Morris, M. R., Hudson, S. E., and Greenberg, S., editors, *CHI Extended Abstracts*, pages 2767–2770. ACM.

[Bramoullé and Rogers, 2009] Bramoullé, Y. and Rogers, B. W. (2009). Diversity and popularity in social networks. Technical report, Discussion paper//Center for Mathematical Studies in Economics and Management Science.

[Bronfenbrenner, 1977] Bronfenbrenner, U. (1977). Toward and experimental ecology of human development. *American Psychologist*, 32(7):513–531.

[Brown, 1986] Brown, F. M. (1986). A commonsense theory of nonmonotonic reasoning. In *8th International Conference on Automated Deduction*, pages 209–228. Springer.

[Buskens and Van de Rijt, 2008] Buskens, V. and Van de Rijt, A. (2008). Dynamics of networks if everyone strives for structural holes1. *American Journal of Sociology*, 114(2):371–407.

[Cairns et al., 1988] Cairns, R. B., Cairns, B. D., Neckerman, H. J., Gest, S. D., and Gariépy, J.-L. (1988). Social networks and aggressive behavior: Peer support or peer rejection? *Developmental psychology*, 24(6):815.

[Carlson et al., 2010] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R., and Mitchell, T. M. (2010). Toward an architecture for never-ending language learning. In *AAAI*.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

[Chklovski and Gil, 2005] Chklovski, T. and Gil, Y. (2005). An analysis of knowledge collected from volunteer contributors. In *AAAI'05: Proceedings of the 20th national conference on Artificial intelligence*, pages 564–570. AAAI Press.

[Cioffi-Revilla, 2010] Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):259–271.

[Cirucci, 2013] Cirucci, A. M. (2013). First person paparazzi: Why social media should be studied more like video games. *Telemat. Inf.*, 30(1):47–59.

[Clauset et al., 2004] Clauset, A., Newman, M. E., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):066111.

[Consolvo et al., 2009] Consolvo, S., McDonald, D. W., and Landay, J. A. (2009). Theory-driven design strategies for technologies that support behavior change in everyday life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 405–414. ACM.

[Cosley et al., 2003] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., and Riedl, J. (2003). Is seeing believing?: how recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 585–592. ACM.

[Cosley et al., 2012] Cosley, D., Sosik, V., Schultz, J., Peesapati, S. T., and Lee, S. (2012). Experiences with designing tools for everyday reminiscing. In *HCI*, volume 27, pages 175–198.

[Crandall, 1988] Crandall, C. S. (1988). Social contagion of binge eating. *Journal of personality and social psychology*, 55(4):588.

[Dabbish et al., 2012] Dabbish, L., Stuart, C., Tsay, J., and Herbsleb, J. (2012). Social coding in github: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1277–1286. ACM.

[Dinakar et al., 2012] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

[Dodge and Coie, 1987] Dodge, K. A. and Coie, J. D. (1987). Social-information-processing factors in reactive and proactive aggression in children's peer groups. *Journal of personality and social psychology*, 53(6):1146.

[Dodge et al., 1986] Dodge, K. A., Pettit, G. S., McClaskey, C. L., and Brown, M. M. (1986). Social competence in children. *Monographs of the society for research in child development*.

[Donath et al., 1999] Donath, J., Karahalios, K., and Viegas, F. (1999). Visualizing conversation. *Journal of Computer-Mediated Communication*, 4(4):0–0.

[Elio, 2002] Elio, R. (2002). Issues in commonsense reasoning. In Elio, R., editor, *Commonsense, Reasoning and Rationality*, pages 3–36. Oxford University Press, New York.

[Elkan and Noto, 2008] Elkan, C. and Noto, K. (2008). Learning classifiers from only positive and unlabeled data. In *Proceedings of the KDD'2008*, pages 213–220.

[Emmelkamp, 2005] Emmelkamp, P. M. (2005). Technological innovations in clinical assessment and psychotherapy. *Psychotherapy and psychosomatics*, 74(6):336–343.

[Ennett et al., 2008] Ennett, S., Faris, R., Hipp, J., Foshee, V., Bauman, K., Hussong, A., and Cai, L. (2008). Peer smoking, other peer attributes, and adolescent cigarette smoking: A social network analysis. *Prevention Science*, 9:88–98.

[Espelage and Holt, 2001] Espelage, D. and Holt, M. (2001). Bullying and victimization during early adolescence: Peer influences and psychosocial correlates. *Journal of Emotional Abuse*, 2:123–142.

[Espelage et al., 2003] Espelage, D., Holt, M., and Henkel, R. (2003). Examination of peer-group contextual effects on aggression during early adolescence. *Child Development*, 74:205–220.

[Espelage and Horne, 2008] Espelage, D. and Horne, A. (2008). School violence and bullying prevention: From research based explanations to empirically based solutions. *Handbook of counseling psychology*, pages 588–606.

[Espelage et al., 2004] Espelage, D., Mebane, S., and Adams, R. (2004). Empathy, caring, and bullying: Toward an understanding of complex associations. In *Bullying in American Schools: A social-ecological perspective on prevention and intervention*, pages 37–61, New Jersey, NJ, USA. Lawrence Erlbaum Associates.

[Friedman and Kahn, 2003] Friedman, B. and Kahn, Jr., P. H. (2003). The human-computer interaction handbook. chapter Human Values, Ethics, and Design, pages 1177–1201. L. Erlbaum Associates Inc., Hillsdale, NJ, USA.

[Garandeau et al., 2011] Garandeau, C. F., Ahn, H.-J., and Rodkin, P. C. (2011). The social status of aggressive students across contexts: The role of classroom status hierarchy, academic achievement, and grade. *Developmental psychology*, 47(6):1699.

[Geelen et al., 2012] Geelen, D., Keyson, D., Boess, S., and Brezet, H. (2012). Exploring the use of a game to stimulate energy saving in households. *Journal of Design Research*, 10(1):102–120.

[Goel et al., 2010] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences*, 107(41):17486–17490.

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18.

[Handcock, 2003] Handcock, M. S. (2003). Statistical models for social networks: Inference and degeneracy. *Dynamic social network modeling and analysis*, 126:229–252.

[Hawley et al., 2007] Hawley, P., Little, T., and Rodkin, P. (2007). *Aggression and Adaptation: The Bright Side to Bad Behavior*. Taylor & Francis.

[Hawley, 2003] Hawley, P. H. (2003). Prosocial and coercive configurations of resource control in early adolescence: A case for the well-adapted machiavellian. *Merrill-Palmer Quarterly*, 49.3:279–309.

[Hogan, 2010] Hogan, B. (2010). The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6):377–386.

[IBM, 2011] IBM (2011). Ibm cityone.

[Isaacs et al., 2013] Isaacs, E., Konrad, A., Walendowski, A., Lennig, T., Hollis, V., and Whittaker, S. (2013). Echoes from the past: how technology mediated reflection improves well-being. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1071–1080, New York, NY, USA. ACM.

[Jackson and Wolinsky, 1996] Jackson, M. O. and Wolinsky, A. (1996). A strategic model of social and economic networks. *Journal of economic theory*, 71(1):44–74.

[Karahalios and Viégas, 2006] Karahalios, K. G. and Viégas, F. B. (2006). Social visualization: Exploring text, audio, and video interaction. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 1667–1670, New York, NY, USA. ACM.

[Keresteš and Milanović, 2006] Keresteš, G. and Milanović, A. (2006). Relations between different types of children's aggressive behavior and sociometric status among peers of the same and opposite gender. *Scandinavian Journal of Psychology*, 47(6):477–483.

[Khatib et al., 2011] Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., Baker, D., and Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47):18949–18953.

[Khovanskaya et al., 2013] Khovanskaya, V., Baumer, E. P., Cosley, D., Voida, S., and Gay, G. (2013). "everybody knows what you're doing": a critical design approach to personal informatics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 3403–3412, New York, NY, USA. ACM.

[Kumar et al., 2010] Kumar, R., Novak, J., and Tomkins, A. (2010). Structure and evolution of online social networks. In *Link mining: models, algorithms, and applications*, pages 337–357. Springer.

[Law and Ahn, 2011] Law, E. and Ahn, L. v. (2011). Human computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 5(3):1–121.

[Lazer et al., 2009] Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721.

[Lenat et al., 1990] Lenat, D. B., Guha, R. V., Pittman, K., Pratt, D., and Shepherd, M. (1990). Cyc: toward programs with common sense. *Commun. ACM*, 33(8):30–49.

[Lenhart et al., 2008] Lenhart, A., Arafeh, S., Smith, A., and Macgill, A. (2008). Writing, technology and teens, pew internet & american life project.

[Leshed et al., 2009] Leshed, G., Perez, D., Hancock, J. T., Cosley, D., Birnholtz, J., Lee, S., McLeod, P. L., and Gay, G. (2009). Visualizing real-time language-based feedback on teamwork behavior in computer-mediated groups. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 537–546. ACM.

[Liben-Nowell and Kleinberg, 2007] Liben-Nowell, D. and Kleinberg, J. (2007). The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031.

[Lieberman et al., 2007] Lieberman, H., Smith, D., and Teeters, A. (2007). Common consensus: a web-based game for collecting commonsense goals. In *ACM Workshop on Common Sense for Intelligent Interfaces*.

[Lifschitz, 1995] Lifschitz, V. (1995). The logic of common sense. *ACM Computing Surveys (CSUR)*, 27(3):343–345.

[Light and McGrath, 2010] Light, B. and McGrath, K. (2010). Ethics and social networking sites: a disclosive analysis of Facebook. *Information Technology & People*, 23:290–311.

[Littlefield-Cook et al., 2005] Littlefield-Cook, J., Cook, G., Berk, L. E., and Bee, H. (2005). *Child development: Principles and perspectives*, volume 55. Allyn and Bacon.

[Liu et al., 2002] Liu, H., Lieberman, H., and Selker, T. (2002). Goose: a goal-oriented search engine with commonsense. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 253–263. Springer.

[Liu and Singh, 2004] Liu, H. and Singh, P. (2004). Conceptneta practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

[L.Jedrzejczyk et al., 2009] L.Jedrzejczyk, B.A.Price, A.K.Bandara, and B.Nuseibeh (2009). I know what you did last summer: risks of location data leakage in mobile and social computing. Technical Report 2009/11.

[Lopes et al., 2013] Lopes, M. C., Fialho, F. A., Cunha, C. J., and Niveiros, S. I. (2013). Business games for leadership development a systematic review. *Simulation & Gaming*, 44(4):523–543.

[Madan et al., 2012] Madan, A., Cebrian, M., Moturu, S., Farrahi, K., and Sandy, A. (2012). Sensing the health state of a community. *IEEE Pervasive Computing*, 11(4):36–45.

[Maloof, 2003] Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. In *ICML 2003 Workshop on learning from imbalanced data sets II*, volume 21, pages 1263–1284.

[Mancilla-Caceres and Amir, 2011] Mancilla-Caceres, J. F. and Amir, E. (2011). Evaluating commonsense knowledge with a computer game. In Campos, P. F., Graham, T. C. N., Jorge, J. A., Nunes, N. J., Palanque, P. A., and Winckler, M., editors, *INTERACT (1)*, volume 6946 of *Lecture Notes in Computer Science*, pages 348–355. Springer.

[Mancilla-Caceres et al., 2013] Mancilla-Caceres, J. F., Amir, E., and Espelage, D. (2013). Peer nominations and its relation to interactions in a computer game. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 38–47. Springer.

[Mancilla-Caceres et al., 2012a] Mancilla-Caceres, J. F., Pu, W., Amir, E., and Espelage, D. (2012a). A computer-in-the-loop approach for detecting bullies in the classroom. In Yang, S. J., Greenberg, A. M., and Endsley, M. R., editors, *SBP*, volume 7227 of *Lecture Notes in Computer Science*, pages 139–146. Springer.

[Mancilla-Caceres et al., 2012b] Mancilla-Caceres, J. F., Pu, W., Amir, E., and Espelage, D. (2012b). Identifying bullies with a computer game. In Hoffmann, J. and Selman, B., editors, *AAAI*. AAAI Press.

[Matuszek et al., 2005] Matuszek, C., Witbrock, M., Kahlert, R. C., Cabral, J., Schneider, D., Shah, P., and Lenat, D. (2005). Searching for common sense: Populating cyc from the web. In *In Proceedings of the Twentieth National Conference on Artificial Intelligence*, pages 1430–1435.

[McCarthy, 1968] McCarthy, J. (1968). Programs with common sense. In *Semantic Information Processing*, pages 403–418. MIT Press.

[McCarthy, 1989] McCarthy, J. (1989). Artificial intelligence, logic and formalizing common sense. In *Philosophical Logic and Artificial Intelligence*, pages 161–190. Springer.

[McCarthy, 1990] McCarthy, J. (1990). Artificial intelligence, logic and formalizing common sense. In *Philosophical Logic and Artificial Intelligence*, pages 161–190. Kluwer Academic.

[McDermott and Doyle, 1980] McDermott, D. and Doyle, J. (1980). Non-monotonic logic i. *Artificial intelligence*, 13(1):41–72.

[McGonigal, 2011] McGonigal, J. (2011). Reality is broken. *Jonathan Cape, London*.

[McPherson et al., 2001] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

[Microsoft, 2012] Microsoft (2012). Worldwide online bullying research. http://www.microsoft.com/security /resources/research.aspx#onlinebullying.

[Monks et al., 2005] Monks, C. P., Smith, P. K., and Swettenham, J. (2005). Psychological correlates of peer victimization in preschool: social cognitive skills, executive function and attachment profiles. *Aggressive Behavior*, 31(6):571–588.

[Motahari et al., 2009] Motahari, S., Ziavras, S., Schuler, R., and Jones, Q. (2009). Identity inference as a privacy risk in computer-mediated communication. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, pages 1–10.

[Naubuzoka, 2009] Naubuzoka, D. (2009). Teacher ratings and peer nominations of bullying and other behaviour of children with and without learning difficulties. *Educational Psychology*, 23(3):307–321.

[Noto et al., 2008] Noto, K., Saier, M., and Elkan, C. (2008). Learning to find relevant biological articles without negative training examples. In *Ai 2008: Advances in Artificial Intelligence*, volume 5360, pages 202–213.

[Ohm, 2009] Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review, Vol. 57, p. 1701, 2010.*

[Pearl, 1988] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Plaisant et al., 1996] Plaisant, C., Milash, B., Rose, A., Widoff, S., and Shneiderman, B. (1996). Lifelines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 221–227. ACM.

[Porter et al., 2004] Porter, S. R., Whitcomb, M. E., and Weitzer, W. H. (2004). Multiple surveys of students and survey fatigue. *New Directions for Institutional Research*, 2004(121):63–73.

[Pu et al., 2012] Pu, W., Amir, E., and Espelage, D. (2012). Approximation partition functions for exponential-family random graph models. *Workshop on Algorithmic and Statistical Approaches for Large Social Networks NIPS2012*.

[Quinn and Bederson, 2011] Quinn, A. J. and Bederson, B. B. (2011). Human computation: a survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1403–1412. ACM.

[Reed et al., 2002] Reed, S. L., Lenat, D. B., et al. (2002). Mapping ontologies into cyc. In *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web*, pages 1–6.

[Reiter, 1980] Reiter, R. (1980). A logic for default reasoning. *Artificial intelligence*, 13(1):81–132.

[Ritterfeld et al., 2009] Ritterfeld, U., Cody, M., and Vorderer, P. (2009). *Serious games: Mechanisms and effects*. Routledge.

[Robertson et al., 2009] Robertson, S., Vojnovic, M., and Weber, I. (2009). Rethinking the esp game. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 3937–3942. ACM.

[Robins et al., 2007] Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2):173 – 191. ¡ce:title¿Special Section: Advances in Exponential Random Graph (p*) Models¡/ce:title¿.

[Scaiano, 2012] Scaiano, M. (2012). Populating a knowledge base from a dictionary. In *Advances in Artificial Intelligence*, pages 392–395. Springer.

[Schlkopf et al., 2001] Schlkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., and Williamson, R. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471.

[Schölkopf et al., 2000] Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural Comput.*, 12:1207–1245.

[Seigne et al., 2007] Seigne, E., Coyne, I., Randall, P., and Parker, J. (2007). Personality traits of bullies as a contributory factor in workplace bullying: An exploratory study. *International Journal of Organization Theory and Behavior*, 10(1):118–132.

[Shahaf and Amir, 2007] Shahaf, D. and Amir, E. (2007). Towards a theory of ai completeness. In *8th International Symposium on Logical Formalizations of Commonsense Reasoning (Commonsense'07)*.

[SigArt, 2009] SigArt (2009). Wikipedia knowledge extractor. Special Interest Group on Artificial Intelligence (SIGArt), Association of Computing Machinery (ACM), University of Illinois at Urbana-Champaign (UIUC).

[Singh et al., 2002] Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. pages 1223–1237. Springer-Verlag.

[Snijders, 2002] Snijders, T. A. B. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3.

[Suler, 2004] Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & behavior*, 7(3):321–326.

[Tang et al., 2011] Tang, J. C., Cebrian, M., Giacobe, N. A., Kim, H.-W., Kim, T., and Wickert, D. B. (2011). Reflecting on the darpa red balloon challenge. *Communications of the ACM*, 54(4):78–85.

[Tausczik and Pennebaker, 2013] Tausczik, Y. R. and Pennebaker, J. W. (2013). Improving teamwork using real-time language feedback. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 459–468, New York, NY, USA. ACM.

[TheWhiteHouse, 2011] TheWhiteHouse (2011). The white house conference on bullying prevention, press release. http://www.whitehouse.gov/the-press-office/2011/03/10/background-white-house-conference-bullying-prevention.

[Varga et al., 2010] Varga, E., Furlan, B., and Milutinovic, V. (2010). Document filter based on extracted concepts. *IPSI Transaction on Internet Research*, 6(1):5–9.

[Viégas et al., 2004] Viégas, F. B., Wattenberg, M., and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM.

[von Ahn, 2013] von Ahn, L. (2013). Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 1–2. ACM.

[von Ahn and Dabbish, 2004] von Ahn, L. and Dabbish, L. (2004). Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA. ACM.

[von Ahn and Dabbish, 2008] von Ahn, L. and Dabbish, L. (2008). Designing games with a purpose. *Commun. ACM*, 51(8):58–67.

[von Ahn et al., 2006] von Ahn, L., Kedia, M., and Blum, M. (2006). Verbosity: a game for collecting common-sense facts. In *In Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems, volume 1 of Games*, pages 75–78. ACM Press.

[Von Ahn et al., 2008] Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. (2008). recaptcha: Human-based character recognition via web security measures. *Science*, 321(5895):1465–1468.

[Wang et al., 2007] Wang, F.-Y., Carley, K. M., Zeng, D., and Mao, W. (2007). Social computing: From social informatics to social intelligence. *Intelligent Systems, IEEE*, 22(2):79–83.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Number 8 in Structural analysis in the social sciences. Cambridge University Press, 1 edition.

[Werbach and Hunter, 2012] Werbach, K. and Hunter, D. (2012). *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton Digital Press.

[West et al., 2009] West, R., Pineau, J., and Precup, D. (2009). Wikispeedia: An online game for inferring semantic distances between concepts. In *IJCAI*, pages 1598–1603.

[Xu et al., 2012] Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *HLT-NAACL*, pages 656–666. The Association for Computational Linguistics.

[Yadati and Narayanam, 2011] Yadati, N. and Narayanam, R. (2011). Game theoretic models for social network analysis. In *Proceedings of the 20th international conference companion on World wide web*, pages 291–292. ACM.

[Yakowitz, 2011] Yakowitz, J. (2011). Tragedy of the data commons. *Harvard Journal of Law and Technology*, 25.

[Zimmer, 2010a] Zimmer, M. (2010a). "but the data is already public": On the ethics of research in facebook. *Ethics and Inf. Technol.*, 12(4):313–325.

[Zimmer, 2010b] Zimmer, M. (2010b). but the data is already public: on the ethics of research in facebook. *Ethics and Information Technology*, 12(4):313–325.